ALGORITHMIC BIAS AND DISCRIMINATION: AN INTERDISCIPLINARY PERSPECTIVE

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

Evan Rose

Spring, 2023

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

S. Travis Elliott, Department of Engineering and Society

INTRODUCTION

Developments in machine learning and artificial intelligence (AI) over the past several years have enabled breakthroughs in many different areas, including medicine, autonomous vehicles, computational biology, image and text generation, and many more. As the successes of AI and machine learning become even more abundant, the technologies will move into increasingly specialized applications, including critical applications.

If AI and machine learning are to continue to serve critical roles in everyday life, then they must be implemented as fair, secure systems. In particular, they must respect the social values of the social context into which they are embedded, and furthermore they must be robust to attacks from adversaries who wish to compromise them. This paper studies the ways in which AI technologies are influenced by, and in turn influence, the social context in which they operate. Specifically, this paper studies the social and technical considerations related to issues of fairness, bias, and discrimination in AI technologies. By analyzing the interplay between these factors, we construct a cohesive network of sociotechnical relationships by which the development of such technologies can be understood.

In this paper, we restrict our discussion primarily to algorithmic systems which rank or score individuals in order to evaluate their fitness for some opportunity, such as college admissions, employment screenings, or loan applications. In particular, this excludes a large class of algorithms which operate towards some other goal, but which nonetheless have a profound impact on life outcomes for different groups of people. Moreover, we mostly focus on machine learning decision-making systems, where the decision-making process relies on identifying statistical patterns in previously existing data. However, while our discussion is tailored specifically for the machine-learning setting, many of the arguments extend naturally to the broader setting (where, for example, hand-crafted algorithms may be used for decision making processes).

The paper is structured as follows. First, we list various definitions of bias from different disciplines, giving special care to distinguish between technical and social notions of bias and discrimination. Secondly, we introduce the Social Construction of Technology (SCOT) sociotechnical framework, which we use to model the sociotechnical relationships between algorithmic decision-making technology and social groups it affects. Next, we discuss a motivating example of algorithmic discrimination in the criminal justice system, and explore different notions of fairness and bias as they relate to the example. Finally, we complete our analysis by formulating the relationships between decision-making technology and the affected social groups.

BACKGROUND

DEFINITIONS OF BIAS

Before attempting to have any meaningful discussion about bias, one first needs to clarify what it means to say a system is biased. The term *bias* has several different meanings depending on the context or discipline in which it is used. In statistics, bias may take on one of several different technical meanings. One common definition describes the tendency of an estimator to consistently deviate from the true value of the predicted quantity in some direction (Wasserman, 2010, sec. 6.3). Another common definition in statistics relates to systematic errors which influence the outcome of an experiment and affect what conclusions can be reliably drawn from it. For example, experimental errors such as sampling bias may cause skewed results in an experiment. Yet another definition appears in machine learning, where bias refers to a parameter which is added to a model’s raw output to obtain a more accurate prediction (Géron, 2019, p. 112). While some of these technical notions of bias do have the potential to cause significant ethical problems, they are not the main focus of our discussion. Instead, we will be primarily concerned with those definitions of bias which relate to the tendencies of a system to behave in morally objectionable ways, especially when the behavior of the system depends highly on the groups of people who interact with it. For the purposes of this paper, (algorithmic) bias refers to consistent tendencies of a system to behave in morally or socially objectionable ways based on the demographic properties of the person affected. This definition is deliberately open-ended, and later more comprehensive normative arguments which determine which systems may be labeled as biased will be explored.

While the technical notions of bias are not the main subject of our discussion, they are not disconnected from the social notions of bias. The presence of certain forms of technical bias—especially those which directly relate to the behavior of a system on a population of individuals, such as sampling bias or estimator bias—can often result in ethically undesirable behavior. In fact, the ethically relevant forms of bias that can be ideally erased from algorithmic systems are intimately connected to the technical methods used to evaluate such systems. Due to the ambiguity when using the term “bias,” this paper will distinguish between the different forms of bias when the intended meaning is not immediately clear; the phrase “technical bias” will refer to those types of statistical bias such as sampling bias or the bias of an estimator, while “social bias” will refer to the ethically concerning behavior of a system which serves as the main focus of this paper.

In machine learning, discrimination is a particularly challenging obstacle to overcome, because it is at odds with the typical goals the tool is designed to achieve. Indeed, in many cases the very purpose of a machine learning system is to discriminate between groups in a technical sense: decision-making systems are often tasked with deciding people who qualify for loans (Khandani et al., 2010), get shortlisted for a job interview (Eastwood, 2020), and who are eligible for insurance (MacCarthy, 2019). In other words, it is insufficient to demand simply that decision-making systems *not* discriminate; rather, it must first be established what distinguishes unethical discrimination from ethical discrimination.

SOCIAL CONSTRUCTION OF TECHNOLOGY

The Social Construction of Technology (SCOT) sociotechnical framework captures interactions between technical and social systems, primarily in order to understand the development of a technology and secondarily to understand the influence a technology has on society (Pinch & Bijker, 1984). Specifically, the framework models interactions between the engineer and social groups in order to understand the effect each social group has on the development of a technological artifact, as well as how the development of the artifact influences social groups. The framework achieves this by identifying the key social groups in a given sociotechnical system, problems associated with a technology those social groups face, and feasible solutions to those problems. The framework is multidirectional by design, and contrasts with linear models applied to the history of technology by emphasizing the many different ways in which a sociotechnical system could develop in response to the social and technical forces within it. The fate of the technology is not presupposed as with other linear models of development, and instead is realized only through a process of variation and selection in which technologies undergo several iterations which then are filtered by their utility to the social groups.

The SCOT model begins with the clear identification of key social groups who are affected by the technology being studied. Social groups are defined as “institutions and organizations . . . as well as organized and unorganized groups of individuals” subject to “the key requirement is that all members of a certain social group share the same set of meanings, attached to a specific art[i]fact” (1984, p. 414). The social groups are central to a solid sociotechnical analysis, as the goal of understanding the development of a technology through a sociotechnical lens hinges on understanding the common meaning understood by each social group interacting with the technology. Indeed, to describe a problem with a technology at all assumes that there is a group for the problem to affect. In particular, the social groups must be defined with respect to the technology being studied, and not in isolation of any specific technology. The common meaning associated with the technology heavily informs the inferences that can be drawn about a technological development.

Once the key social groups have been introduced, the development of a technology can be framed with respect to the problems existing between the social groups and the technology. The separation of the social groups allows room for *interpretive flexibility*: namely, the idea that a technology can have different meanings depending on the person or group using or affected by it. By allowing each social group a distinct interpretation of a technology, and by selecting social groups such that the members in a group share a particular set of meanings, SCOT achieves a multidirectional model in which different groups affect and are affected by the development of a technology according to the relevant relationships between them.

Pinch & Bijker also make the following observation regarding the SCOT framework: once the technical meanings and requirements among the different key social groups are established, “various solutions for these conflicts and problems are possible—not only technological, but also judicial, or even moral" (1984, p. 416). Just as the social groups influence the development of a technology through their combined goals and technical requirements, the development and introduction of a new technology into society has the potential to alter social behaviors as well. Thus, while strictly technical solutions may be sufficient to address issues of algorithmic discrimination, it is also possible that the development of such technologies may induce social or moral shifts as a means of resolving these problems.

CASE STUDY: COMPAS RECIDIVISM

Algorithmic risk assessment is a process which takes information about a scenario as input and produces a score reflecting the risk associated with that scenario as output. For example, a loan application screening pipeline could use a risk assessment tool to predict the risk of a borrower defaulting on a loan, providing valuable information for the lender when evaluating loan proposals from applicants (Beshr, 2021; Quinn, 2021). In the criminal justice system, risk assessment instruments are used to predict risks related to defendants and criminals, such as the risk that a defendant fails to appear in court or the risk that a convicted criminal will commit a violent crime in the future (Chohlas-Wood, 2020, para. 2).

Algorithmic risk assessment tools have caused controversy in the past. One such tool, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), is used in the criminal justice system to predict risk scores for various scenarios (Northpointe Inc., 2015). The tool accomplishes this by analyzing data about a defendant's criminal history, demographic information, and other relevant factors, to assist in making decisions about pre-trial release, bond amounts, probation, and parole. One of the uses of the COMPAS tool is to predict recidivism: a relapse into criminal activity after the release of a convicted criminal.

In a recent audit of COMPAS, Angwin et al. argued that the tool exhibited strong social bias against black defendants (2016). The basic argument for the claim was rooted in a statistical analysis that demonstrated a higher false positive rate for black defendants, as well as several examples examining the tool's behavior across different defendants whose recidivism outcome was known. ProPublica found that black defendants who received a high risk score from the tool had a lower overall recidivism rate than white defendants who received a similarly high score. In contrast, white defendants who received low risk scores were more likely to recidivate than black defendants who received similarly low scores. These findings led ProPublica to conclude that COMPAS was biased against black defendants due to the way it fails for different racial groups.

The story prompted lots of discussion surrounding algorithmic fairness and the role of algorithmic decision-making in the criminal justice system. However, some others have written objections to the statistical rigor of Angwin et al.'s investigation (Dieterich et al., 2016; Gong, 2016).

MEASURING ALGORITHMIC FAIRNESS

To understand and contextualize the main arguments made by and against ProPublica's accusation, it is important to understand the different ways in which an algorithm (specifically, one created with the purpose of assigning some score to individuals) can be evaluated for algorithmic bias. For this purpose, we will refer to the three formal fairness criteria identified by Barocas et al. (2019). While other authors have proposed many alternative fairness criteria (Chouldechova, 2017; Dwork et al., 2011; Hardt et al., 2016), the three proposed by Barocas et al. reflect the main goals of previous work and absorb many previously studied criteria. Each fairness criterion captures some ethical ideal related to the way sensitive data should be treated in decision-making systems such as the recidivism tools used in the criminal justice system.

Each fairness criterion can be described as a statistical relationship between three random variables. The first variable is the *sensitive attribute*, which represents a personal trait. In the case of the COMPAS tool, the protected attribute is race; more generally, protected attributes may include sex, gender, sexuality, religion, or other traits which may serve as the basis of morally objectionable discrimination. The second variable is the *target variable*, which is the property the algorithmic system is tasked with predicting for a given individual. In the case of the COMPAS tool, the target variable is a criminal's chance of recidivism; other common objectives include loan default probabilities and potential employee utility (Eastwood, 2020; Khandani et al., 2010). The third variable is the *score*: the system's prediction for the value of the target variable given an individual.

The first fairness criterion identified by Barocas et al. is called *independence*, and states that the score should be (statistically) independent of the protected attribute. Informally, this requirement states that the score should not depend on the protected attribute; that is, knowing the protected attribute for an individual should not affect the likelihood of that individual receiving any particular score. Independence is most desirable when it reflects a fundamental belief about the relationship between two variables; namely, if one already holds the belief that the protected attribute is unrelated to the target variable. Finally, independence may also reflect a desire to instill *equitable opportunity* across a diverse population: for example, if a college admissions screening system tends to reject applicants from low-income socioeconomic backgrounds (even if it is *statistically* justified in doing so, since low-income students may face additional challenges making success in college less likely), a morally defensible position is that the system should distribute *opportunity* in a way which gives disadvantaged groups a chance at catching up—indeed, this forms one of the basic arguments for affirmative action in education and employment.

The second fairness criterion is called *separation*, and states that the score should be independent of the protected attribute, *conditioned on the target variable*. Informally, separation requires that a group of individuals who have equal target variables (such as credit score) should not receive different scores based on the protected attribute. In contrast to independence, separation allows the predictor to correlate with the target variable insofar as the protected attribute genuinely relates to the target, but no further. For example, car insurance premiums tend to be higher for young drivers because they are more likely to be involved in accidents (Coleman, 2022; Deventer, 2023). If the higher premiums were not based on this statistical reality, the different treatment on the basis of age would constitute morally objectionable behavior. Thus, a risk prediction tool that rates young drivers at higher risk of accidents (insofar as they *are* at higher risk) satisfies separation, but a tool which inordinately rates young drivers out of accordance with their actual risk violates separation. Even in situations where separation is desirable, measuring separating can be difficult to practically enforce in cases where the target variable is fundamentally unknown and only visible through proxy variables.

The third fairness criterion is called *sufficiency*, and states that the target variable should be independent of the protected attribute, *conditioned on the score*. Informally, this criterion requires that the score eliminates any predictive advantage to possessing the protected attribute; that is, the score makes knowing the protected attribute irrelevant for the purposes of a decision. For example, a medical test that predicts a patient’s reaction to a treatment should have the same predictive power for all groups, regardless of an attribute like race. If the interpretation of a high risk score differs substantially depending on the racial background of the patient receiving it, then the test violates sufficiency. On the other hand, if a high test score supplies the same predictive information to all patients, regardless of race, then the test satisfies sufficiency.

A fundamental result involving the three fairness criteria is that, within any particular decision-making system, each pair of fairness criteria is mutually exclusive. That is, a predictor cannot simultaneously satisfy both independence and separation, independence and sufficiency, or separation and sufficiency. The example of affirmative action is again a useful illustration. In establishing independence with the goal of distributing opportunity fairly across different groups, a predictor is forced to dispense with separation (since now the score is presumably not completely determined by an applicant's true chances of success) and with sufficiency (since an applicant with a low chance of success will receive a higher score if the applicant is from a disadvantaged group). This tradeoff between fairness and predictive power applies to all decision-making systems and highlights a fundamental tension in creating fair algorithms: not only does achieving fairness require reducing effectiveness, but moreover we must choose what kind of fairness is important to optimize within any particular setting.

OBJECTIONS TO PROPUBLICA

We will now discuss the main objections to ProPublica's accusation against the COMPAS tool, by applying the fairness criteria from Barocas et al. As Data Scientist Abe Gong discusses in an evaluation of ProPublica's investigation, the metrics used by ProPublica to identify discrimination by the COMPAS tool are not among those typically used to measure fairness in algorithms. The main technical argument made by ProPublica hinges on the fact that COMPAS recidivism scores produce a higher false positive rate on black individuals as compared to white individuals, resulting in a model which “made mistakes with black and white defendants at roughly the same rate but in very different ways” (2016). The main objection to this argument is that false positive rate fails to consider the base rate of recidivism between groups (Gong, 2016).

On the other hand, Gong demonstrated that the COMPAS tool satisfies sufficiency: that is, the scores produced by the COMPAS tool correspond consistently with likelihood of recidivism, and moreover race does not affect the accuracy of these scores. By this standard, the tool does not exhibit technical bias against blacks, since individuals from different racial groups are classified consistently in accordance with their true likelihood of recidivism. Crucially, satisfying other measures of fairness (such as the error rate parity as demanded by ProPublica) necessarily requires deviating from this standard. Thus, a fair tool by ProPublica’s standards must produce less predictive scores on certain groups of individuals on the basis of a protected attribute like race.

A BROADER ETHICAL PERSPECTIVE

Regardless of whether or not the precise claims made by ProPublica withstand careful scrutiny, the discussion about the fairness of the COMPAS tool raises deep questions about the role of AI in critical applications such as recidivism prediction. For example, a deeper and perhaps more critical question to ask in this scenario is *why* blacks experience higher base recidivism rates than whites. Answering this question requires a comprehensive inspection of social, legal, and technical systems that extend beyond a strictly technical evaluation of currently implemented tools.

Another issue that may be raised is that recidivism is actually measured in practice with a proxy variable: results from real-world policing. Data approximated by arrest rate is potentially skewed due to historical social biases, and so may have direct implications for a system trained on it. More broadly, an ethical argument could be made that the COMPAS tool implements a form of predictive policing, rendering it categorically unethical, regardless of its technical capabilities. Such fundamental difficulties are critical to examine carefully, as they challenge the legitimacy of the goals of the method on ethical grounds, rather than a particular implementation of the method on technical ones. Gong, despite rejecting the conclusions made by ProPublica, acknowledges these more fundamental issues eloquently: "Powerful algorithms can be harmful and unfair, even when they're unbiased in a strictly technical sense" (2016).

GUIDELINES FOR ASSESSING ALGORITHMIC FAIRNESS

The COMPAS case study suggests a few general guidelines for using AI tools in high-stakes applications. Firstly, the evaluation of decision-making systems for fairness should be aware of the social context in which they are deployed. The COMPAS case study illustrates that any arguments regarding the fairness of a system rely on value-laden data and methods. The choice to focus on race, or any protected attribute for that matter, reflects a legacy of social injustice deeply ingrained in the social fabric, and cannot be determined without such a rich social context. Algorithmic decision-making systems operate within a complex environment that includes law, society, ethics, and business, and arguments about such systems must accordingly account for the broader social, legal, and economic implications of their development and implementation.

While we do not provide similarly detailed analyses for other reports of algorithmic discrimination, it is important to understand what fundamental technical limitations do currently exist and that they are intrinsic to every instance of algorithmic bias. The need for certain performance tradeoffs affects all algorithmic decision systems operating in a social environment, and a thorough conversation could be had about any further examples discussed in this paper. However, similarly detailed discussion of additional reports of algorithmic bias are beyond the scope of this paper, and we will proceed by acknowledging by default the potential limitations of other reports of algorithmic bias.

ALGORITHMIC DECISION-MAKING AS A SOCIAL CONSTRUCTION OF TECHNOLOGY

For the remainder of this paper, we will frame the development of algorithmic decision-making systems as a sociotechnical system by using the SCOT framework. In particular, we identify key social groups involved in the sociotechnical system as well as the way the groups interact with, affect, and are affected by the technology. Then, we discuss the main problems each group faces with respect to the technology. Finally, we discuss a possible set of solutions to the proposed problems.

SOCIAL GROUPS FOR ALGORITHMIC DISCRIMINATION

To apply the SCOT sociotechnical framework to algorithmic fairness, we begin by identifying the key social groups affected by the technology. The full range of social groups affected by algorithmic decision-making systems is extensive, and it is beyond the scope of this paper to discuss all of them. We instead select a few highly salient social groups which reflect many of the broader challenges of achieving algorithmic fairness, with a focus on those groups which have been disproportionately affected by issues of algorithmic fairness as well as those which play a key role in shaping the development of algorithmic decision-making technology. In addition, we identify at least one problem or objective relating each social group to the technology—thereby justifying the formation of the group for the purposes of a sociotechnical analysis—as well as some possibilities for how that problem or objective affects the overall sociotechnical system.

Racial and ethnic minorities serve as a key social group due to their tendency to be underrepresented in the algorithmic design process. We have already discussed one way in which racial groups are affected by algorithmic bias with the COMPAS case study. In the past, datasets used to train decision-making systems have failed to include sufficient data from minority groups, especially racial and ethnic groups (Buolamwini & Gebru, 2018; Wen et al., 2022). As David Wen et al. showed in their review of publicly available skin cancer datasets, there is a "substantial under-representation of darker skin types" (Wen et al., 2022, p. e64). In the few datasets that did report data on ethnicity, "no images were from individuals with an African, Afro-Caribbean, or South Asian background" (2022, p. e71). These limitations are important because they could result in cancer detection models which perform worse on underrepresented groups, since some skin cancers manifest differently depending on skin color (2022, p. e71). To address the data underrepresentation issue directly, data collection processes may need to explicitly consider the diversity of their datasets. Already, some organizations are taking steps to produce more inclusive datasets which achieve better representation of minority groups (Doshi, 2018). Additionally, transparent bias audits such as the one conducted by ProPublica against

COMPAS can counter the propagation of discriminatory practices by giving a voice to the adversely affected social group.

Women have also been found to be disproportionately affected by algorithmic bias in domains such as employment and credit decisions (Datta et al., 2015; Knight, 2019). In 2019, the algorithm used to determine credit lines for the new Apple Card displayed an acute tendency to offer women substantially smaller credit lines than men, even when the two card owners were part of the same family and shared similar credit backgrounds (Knight, 2019). One study conducted by researchers at Carnegie Mellon concluded that Google ad services were more likely to offer high-paying job ads to men than women who otherwise had identical profiles (Datta et al., 2015). And Amazon was forced to discard an employment screening model after the system repeatedly downgraded the scores of resumes that contained the word "women" (Dastin, 2018). These examples demonstrate the importance of including women as a social group in a sociotechnical analysis of AI technologies, as they face distinct shared challenges with the technology. Similarly to racial groups, possible resolutions to the challenges woman face with algorithmic discrimination may include methods for producing more inclusive datasets and bias audits.

Automated systems are a prime target for cyberattacks, making nefarious actors a unique yet important social group to consider in the presence of algorithmic decision-making systems. In stark contrast to most other key social groups, nefarious actors do not have an interest in using the technology for its intended purpose. Instead, they aim to exploit the technology towards some other objective, often one which directly contradicts the original purpose of the technology and which has serious ethical implications. For example, an adversary may attempt to directly influence the system's behavior by introducing fake data into the training process (Shafahi et al., 2018). Such attacks have pressing implications for algorithmic fairness because adversaries can focus their attacks to be especially effective against specific groups of individuals (Jagielski et al., 2021), and moreover the strength of an attack may even depend on the specific demographic properties of the groups being targeted (Rose et al., 2022). Cyberattacks pose an interesting space of possibilities for the development of decision-making technology. If an algorithmic system is vulnerable to manipulation by an adversary, harm may be done to different social groups even if the system's developer has taken precautions to ensure fairness. A currently unexplored research direction is to study the interplay between existing defenses and fairness, as existing defenses may impact the fairness of the resulting system.

Government agencies, especially those with legislative duties, are often tasked with evaluating and regulating emerging technologies, and in doing so form a key social group. One major task of government entities is to crystalize and codify widely accepted ethical standards into law. Historically, the U.S. government has been responsible for identifying "protected classes": groups of people sharing a common characteristic and who have historically been discriminated against on the basis of that characteristic. These protected characteristics include race (Civil Rights Act, 1964), sex (Civil Rights Act, 1964; Equal Pay Act, 1963), religion (Civil Rights Act, 1964), national origin (Civil Rights Act, 1964), immigration status (Immigration Reform and Control Act, 1986), veteran status (Uniformed Services Employment and Reemployment Rights Act, 1994), and disability status (Americans with Disabilities Act, 1990; Rehabilitation Act, 1973), among other things. Notably, the protected characteristics identified by law are classified so due to historical instances of discrimination based on those characteristics. In this way discrimination law tends to be reflective rather than anticipatory, in the sense that it does not try to account for possible future bases for discrimination, deferring instead to widespread historical social patterns. Combined with the rapid growth and integration of AI technologies, legal entities face a distinct challenge when attempting to codify rules about algorithmic discrimination. Because it relies on identifying preexisting patterns of behavior, discrimination law evolves too slowly to respond to fast-paced technical and social changes (Ferrer et al., 2020, p. 76). Despite this, however, government agencies play an important role in codifying precisely what constitutes unethical discrimination and regulating the development of impactful algorithmic systems.

Businesses, and more precisely business executives, also form an important social group, as they are the ones who typically employ algorithmic decision-making systems in a capacity which actually affects the other social groups. Businesses use decision-making systems in applications including employment screening (Dastin, 2018), targeted advertising (Datta et al., 2015), loan qualification screening (Quinn, 2021), insurance qualification (MacCarthy, 2019), customer service (Bernazzani, 2022), and many more. Businesses typically have several motivations for using algorithmic decision systems, including increased efficiency, lower costs, and improved accuracy in decision-making. Interestingly, one of the main reasons historically for using algorithmic decision-making systems is because they were perceived as objective decision-makers, especially as opposed to the humans who were previously responsible for carrying out the same duties. In the future, businesses may be responsible for establishing clear and responsible policies related to algorithmic decision-making systems. An example of organizational policy in this direction can be seen from OpenAI, the developer of the GPT sequence of language models including, most famously, ChatGPT. On February 14, 2023, OpenAI published a blog post discussing their recent technical accomplishments as well as some of their goals and standards for developing fair and ethical AI systems (OpenAI, 2023). In this post, the company outlines some of their beliefs and guidelines for building AI technologies:

> We believe that AI should be a useful tool for individual people, and this customizable by each user up to limits defined by society. Therefore, we are developing an upgrade to ChatGPT to allow users to easily customize its behavior.
>
> This will mean allowing system outputs that other people (ourselves included) may strongly disagree with. Striking the right balance here will be challenging—taking customization to the extreme would risk enabling malicious uses of our technology and sycophantic Ais that mindlessly amplify people's existing beliefs.
>
> . . . One way to avoid undue concentration of power is to give people who use or are affected by systems like ChatGPT the ability to influence those systems' rules. We believe that many decisions about our defaults and hard bounds should be made collectively, and while practical implementation is a challenge, we aim to include as many perspectives as possible. (OpenAI, 2023)

Only time will tell whether OpenAI will follow through on their commitment to produce socially acceptable AI technologies that benefit all of humanity. In the meantime, the technological landscape will continue to change as a result of social forces, and so too will society find ways to adjust to the introduction of such powerful technologies.

CONCLUSION

Recent growth in the popularity of AI and machine learning places a distinct pressure on those developing AI technologies. Not only must developers conscientiously craft their algorithms to be mindful of the social contexts in which they operate, but they must also prepare to reinforce their algorithms against nefarious actors who may wish to override critical systems. On the other hand, society should realize the ways in which it can affect and is affected by the development of AI technologies. As ethical, social, and security concerns continue to surround AI, it will be the social forces which ultimately determine the future of AI technology.

REFERENCES

Americans with Disabilities Act, Pub. L. No. 101–336 (1990).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Bernazzani, S. (2022, March 25). *18 Examples of Chatbots for Customer Service (& How You Should Be Using Them)*. https://blog.hubspot.com/service/customer-service-chatbots

Beshr, S. (2021, July 21). *A Machine Learning Approach To Credit Risk Assessment*. Medium. https://towardsdatascience.com/a-machine-learning-approach-to-credit-risk-assessment-ba8eda1cd11f

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Chohlas-Wood, A. (2020, June 19). Understanding risk assessment instruments in criminal justice. *Brookings*. https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/

Chouldechova, A. (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments* (arXiv:1703.00056). arXiv. https://doi.org/10.48550/arXiv.1703.00056

Civil Rights Act, Pub. L. No. 88–352 (1964).

Coleman, S. (2022, May 23). *Teen Driving Facts and Statistics 2022*. Bankrate. https://www.bankrate.com/insurance/car/teen-driver-facts-and-statistics/

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Datta, A., Tschantz, M. C., & Datta, A. (2015). *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination* (arXiv:1408.6491). arXiv. https://doi.org/10.48550/arXiv.1408.6491

Deventer, C. (2023, March 2). *Auto Insurance Rates by Age in 2023*. Bankrate. https://www.bankrate.com/insurance/car/auto-insurance-rates-by-age/

Dieterich, W., Christina, M., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Doshi, T. (2018, September 6). *Introducing the Inclusive Images Competition*. https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). *Fairness Through Awareness* (arXiv:1104.3913). arXiv. https://doi.org/10.48550/arXiv.1104.3913

Eastwood, B. (2020, August 30). *Exploration-based algorithms can improve hiring quality and diversity*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/exploration-based-algorithms-can-improve-hiring-quality-and-diversity

Equal Pay Act, Pub. L. No. 88–38 (1963).

Ferrer, X., van Nuenen, T., Such, J., Cote, M., & Criado, N. (2020). *Bias and Discrimination in AI: A cross-disciplinary perspective*.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow* (2nd ed.). O'Reilley Media, Inc.

Gong, A. (2016, August 3). Ethics for powerful algorithms (1 of 4). *Medium*. https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning* (arXiv:1610.02413). arXiv. https://doi.org/10.48550/arXiv.1610.02413

Immigration Reform and Control Act, Pub. L. No. 99–603 (1986).

Jagielski, M., Severi, G., Harger, N. P., & Oprea, A. (2021). *Subpopulation Data Poisoning Attacks* (arXiv:2006.14026). arXiv. http://arxiv.org/abs/2006.14026

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Knight, W. (2019, November 19). The Apple Card Didn't "See" Gender—And That's the Problem. *Wired*. https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

MacCarthy, M. (2019, December 6). Fairness in algorithmic decision-making. *Brookings*. https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/

Northpointe Inc. (2015). *Practitioner's Guide to COMPAS Core*. Northpointe Inc.

OpenAI. (2023, February 16). *How should AI systems behave, and who should decide?* Openai.Com. https://openai.com/blog/how-should-ai-systems-behave#OpenAI

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Quinn, M. (2021, November 11). *Machine Learning in Loan Risk Analysis*. https://www.bluegranite.com/blog/machine-learning-in-loan-risk-analysis

Rehabilitation Act, Pub. L. No. 93–112 (1973).

Rose, E., Suya, F., & Evans, D. (2022, September 21). *Poisoning Attacks and Subpopulation Susceptibility*. 5th Workshop on Visualization for AI Explainability. https://uvasrg.github.io/poisoning

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018). *Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks* (arXiv:1804.00792). arXiv. https://doi.org/10.48550/arXiv.1804.00792

Uniformed Services Employment and Reemployment Rights Act, Pub. L. No. 103–353 (1994).

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*.

Wen, D., Khan, S. M., Ji Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X., & Matin, R. N. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, *4*(1), e64–e74. https://doi.org/10.1016/S2589-7500(21)00252-1