The Impact of Cognitive Bias and Algorithmic Formalism in Enabling the Creation of Discriminatory AI Technologies

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

Aishwarya Gavili

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

Hannah Rogers, Department of Engineering and Society

Abstract

From serving as authentication for personal devices to auto completing searches in Google, automated text and image recognition systems are the future of technology. However, they have demonstrated flawed behaviors in their applications with regards to prejudice towards minority groups. For instance, women and people with darker skin complexions tend to be misidentified in facial recognition systems by either being completely unrecognized or by being labeled with derogatory or outdated terms. Evidently, rule-based machine learning algorithms, specifically those utilized by image and text classification systems, enable racial and gender inequity from the cyclical nature of cognitive bias injection they undergo. Such inequity also results from the adherence to algorithmic formalism or formalist thinking instead of algorithmic realism, which is proposed as a solution to bias. The STS SCOT framework will be used to drive this paper's analysis to understand how different stakeholders involved in the algorithm development process interpret and facilitate the identification of a biased algorithm's "social peak" for which it is deemed acceptable to deploy.

Introduction

Machine learning models, particularly rule-based models, have been notorious for being influenced by the human conscience through the embedding of algorithmic and interpretive biases before and after deployment. In turn, the applications of such models have entailed prejudice towards women and racial minorities, posing moral and ethical concerns of rule-based machine learning in society. In fact, a popular AI classification system called ImageNet Roulette was forced to remove 600,000 images from its online database as more than half of the images exuded racial bias in how they were labeled or categorized, highlighting the gravity of the issue (Solly, 2019). The root of the issue stems from the cyclical nature of bias injection in rule-based models. From the start, humans and algorithms engage in an interactive process to determine how an algorithm will operate to best fit the needs and goals of a human entity. Then, algorithms receive sampled data, oftentimes labeled and unchecked, at the discretion of a human entity. Lastly, a human reacts to the output of a model and makes decisions based on biased information, which will probably be consumed by algorithms later, contributing to the endless cycle. In order to understand how human and algorithmic interactions result in algorithmic biases, how cognitive biases affect the interpretation of model outputs, and how to assess and debias models, this thesis will analyze different image and text classifier systems and related technologies. This thesis will perform its analysis using the SCOT framework to understand how the social context of when a machine learning model was developed facilitates the identification of its "social peak" for which it is deemed acceptable for deployment. First, it is important to understand the different connotations of bias employed in this paper. Cognitive bias can be defined as 'a systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments" (Kliegr et al., 2021). Interpretive biases are cognitive biases that are employed in interpreting scenarios or events. Finally, algorithmic biases are systematic errors in a computer program that create unfair outcomes in its output (Wikipedia contributors, 2021). The implications of algorithmic formalism and algorithmic realism will also further be explored.

Algorithmic Formalism

Cognitive biases are weaved into both algorithmic and interpretive biases through algorithmic formalism or formalist thinking. Algorithmic formalism refers to when algorithms adhere to certain rules and require 'explicit mathematical articulation of inputs, outputs, and goals', without really taking into account the complexity of the real world in their applications. In other words, through formalism there is an increased level of objectivity and neutrality (Green & Viljoen). Here, a paradox surfaces, as an increased level of objectivity is ultimately the reason behind bias or subjectivity in machine learning algorithms.

Objectivity can be defined as being based on facts and not being influenced by personal beliefs (Cambridge Dictionary). In science and technology, it has served as the foundation for many discoveries because of how its use in the scientific method has prevented bias to seep into the production of scientific knowledge (Stemwedel, 2013). However, historically, social power has tended to convey the idea of objectivity, making it a contingent social product. More specifically, in the 19th century, quantification and standardization emerged as a way to assign public accountability with the rise of constitutional and democratic governments. Additionally, new professions and new societal arrangements led individuals occupying these roles to use their scientific knowledge to legitimize their approaches and gain status (Hagendijk, 1999). Moreover, quantification and objectivity has remained as a way for people of power to legitimize their authority and decisions in modern times. This has especially been evident in the realm of Big Tech, where algorithms are developed as products of objectivity to ultimately benefit the executives of the companies who deploy them; they further elevate their white, cisgender male privilege.

To begin with, algorithms emerge from technical considerations such as data availability and model accuracy (Green & Viljoen). When looking at data availability, a majority of the image and text data available on the internet has been an accumulation of files and records throughout history. Consequently, such data also reflects the bigoted sentiments present throughout history. For instance, the ImageNet database mentioned earlier uses Princeton's WordNet to map categories to more than 14 million images. Most of the terms that constitute Princeton's WordNet are built off pre-1972 Library of Congress taxonomies, which contain racist and misogynistic terms (Caliskan, 2022). Additionally, online sourced datasets, which are the most accessible, shift towards Western media default for demographic representation, skewing demographic distributions in facial recognition data sets (Raji & Fried, 2021). As a result, images with women of color are more likely to be labeled related to physical appearance compared to white men and are recognized at substantially lower rates compared to white men (Schwemmer et al., 2020). Similarly, predictive policing is upheld by common definitions of crime present in data, which stem from the classist and racist history of the United States. Consequently, predictive policing algorithms such as the COMPAS algorithm demonstrate an association between black men and criminality, further fueling an unjust criminal justice system (Green & Viljoen). Evidently, the objectivity present in using the most available data in rule-based algorithms simply results in the reproduction of existing social conditions, leading to the optimization of an unjust status quo instead of challenging it (Green & Viljoen). However, it is vital to recognize that the most available data is also data that is free, making it a mode of profit for big tech executives who use it to train their algorithms. In other words, big tech executives use freely bought data and sell it through the discriminatory algorithms they deploy, allowing them to make money without losing any money. And most of the time, such data is obtained from non-consensual means. In fact, in the case of facial recognition predictive policing, the open data sets used are based on pre-conviction mugshots of individuals who may not end up with a criminal record or of individuals who are no longer alive (Healey, 2020). Similarly, images used by ImageNet are collected from search engines like Google, which appropriate people's selfies and vacation pictures without their knowledge (Crawford & Paglen). Not only are minority groups facing the discrimination imposed by biased algorithms, but they are also victims of data privacy violations in relation to the data being mined for these biased algorithms. By leveraging the SCOT framework, there is a clear distinction in the understanding of the underlying mechanisms of the data being used to train AI technologies between the producers and consumers of the technologies. From the start, producers or big tech executives pursue the financial incentives that come with using the most available data. On the contrary, consumers are simply clueless about the data being used, which they are unknowingly contributing to. This difference in interpretation of the inner workings and purpose of AI algorithms between different social groups is paralleled with the quantifiable performance metrics used to evaluate them.

Model accuracy in addition to other mathematical performance metrics such as efficiency largely drive the evaluation of algorithms in algorithmic formalism, excluding social considerations. This phenomena has been coined as 'datafication' in STS literature, which has resulted in a culture that is 'shaped and populated with numbers, where trust and interest in anything that cannot be quantified diminishes' (Beer, 2016). In fact, this culture is present in Big Tech companies where financial incentives such as profit create discriminatory algorithms or products. By leveraging the SCOT framework, the social groups that mainly care about quantifiable outcomes are the predominantly white male executives that make the final decisions on whether an algorithm or product is deployed. For instance, in 2016 Amazon excluded certain neighborhoods from its same-day Prime delivery system as it saw a positive correlation between doing so and its profitability model. In turn, using numeric performance metrics such as revenue resulted in the exclusion of poor, predominantly African American neighborhoods (Ingold & Soper, 2016). Similarly, prior to 2020, Amazon sold its facial recognition product called Amazon Rekognition to police departments and generated large profits, despite being aware of catering to a criminal justice system that disproportionately targets African Americans (Hamilton, 2021). At the end of the day, executives like Jeff Bezos who have deployed discriminatory technologies, view such technologies as mere revenue generators and therfore will only take into account high profit as an indicator of a successful algorithm. Another relevant social group that also evaluates algorithms using quantifiable performance metrics is the computer scientists that develop the backend algorithms for these discriminatory technologies. For them, they work to get compensated. And they are compensated based on the quality of their work, which is positively correlated with high algorithm accuracy. For this reason, computer scientists don't consider the social responsibility they have with their work and will deem their developed algorithms as acceptable if they merely have high performance. While gauging every social situation an algorithm could handle is beyond the scope of their work, computer scientists should reason more thoroughly about when certain metrics should be considered or ignored, especially numeric performance metrics.

Algorithmic Realism and Debiasing

An effective approach to debiasing is rooted in training humans to improve their statistical reasoning, mainly through algorithmic realism. Algorithmic realism involves a porous and contextual approach to evaluating algorithms by highlighting the need for additional modes of analysis (Green & Viljoen). More specifically, it focuses on introducing a level of subjectivity by formulating research questions and selecting methodologies and evaluation metrics that focus on social outcomes and not algorithm quality. Proposed realism solutions include developing bias impact statements, engaging with stakeholders, and unbiasing biased datasets (Lee et al., 2022).

A bias impact statement is a template of questions that can guide developers through the design, implementation, and monitoring phases of an algorithm. Some of these questions include, “What will the automated decision do?”, “How will potential bias be detected?”, “What are the operator incentives?”, “How are other stakeholders being engaged?”, and “Has diversity been considered in the design and execution?”. Andrew Selbst, a law professor at the University of California who specializes in AI and the law, stated that “With an impact assessment, you're being very transparent about how you as a company are approaching the fairness question,” (Hao, 2020). This is very important because ‘fairness’ can have different definitions depending on the people and contexts employing it. In fact, some institutions such as New York University’s AI Now Institute have already employed a model framework that federal entities use to create AIAs or algorithmic impact assessments, which gauge the potential negative effects of an algorithm and emulate the questions present in a bias impact statement (Reisman et al., 2018).

The engagement of stakeholders is also a vital part of taking a realism approach to constructing algorithms. It entails getting users involved in the process of developing and engaging with the algorithms early on, ultimately leading to improved user experiences. Afterall, as Microsoft's senior principal researcher Rich Caruana said "Tech succeeds when users understand the product better than its designers" (Lee et al., 2022). For instance, if external sources such as advisory councils and civil society organizations (ex. NGOs) could help programmers decide on the inputs and outputs of the automated decisions in algorithms, bias can be avoided more easily early on. However, this approach raises the question of who to ask and how to locate the people participating in these groups, hindering its feasibility. In fact, entities similar to advisory councils have been employed and have failed at companies like Google in vetting every stage of the design process. In Google's case, Google's AI ethics board collapsed when certain board members had radical views and were collectively unable to reflect broader society's values (Samuel, 2022). Julia Stoyanovich, director of the NYU Center for Responsible AI, responded to Google's debacle and emphasized the need for public participation and meaningful public input instead of input from one group of people or a group of professionals. Accordingly, public participation through citizen assemblies and mini-publics would be a practical alternative, given a sufficient public understanding of AI (Data Justice Lab). Currently, such an understanding doesn't exist, however with increased accessible AI education, the public could make informed decisions, democratically, about these AI technologies.

Additionally, diversity-in-design should be taken into consideration. More specifically, developers of algorithms should also consider the role of diversity within work teams to emphasize cultural sensitivity, instead of just considering the role of diversity in the training data used. For instance, according to a 2011 study done by the National Institute of Standards and

Technologies (Nist), facial recognition software demonstrated higher accuracy on Asian faces when it was created by Asian firms, illustrating how who makes the software strongly influences how it works (Breland, 2017). Therefore, if there is greater minority representation in developer teams, data and algorithm design considerations pertaining to minorities will not be overlooked.

Lastly, unbiasing biased data sets is a realist approach that can be taken to remove racial and gender inequities from the input data of algorithms. However, it is imperative that data isn't being completely manipulated to receive desired results, but is instead being modified to target the issues previously observed in older existing datasets. An instance of what not to do involves eliminating all known social group associations in word embeddings, which would lead to inaccurate representations of the real world. In turn, incorrect occupational gender statistics would be reflected in many natural language processing or text classification system models (Caliskan, 2022).

In contrast, there have been multiple datasets that have recently been adjusted and developed in a more acceptable manner including the WinoBias dataset, Pilot Parliaments Benchmark dataset, and Diversity in Faces (DiF) dataset. The WinoBias dataset follows a winograd format with 40 occupations referenced by different human pronouns, and it has been used to certify whether a system contained gender biases (Mehrabi et al., 2021). The Parliaments Benchmark dataset contains images of 1270 individuals from European and African national parliaments and has been applauded for its gender and racial balance as well as its diversity. Lastly, the DiF dataset contains one million annotations of face images, containing diverse facial features (ex. Craniofacial distances, skin color, facial symmetry), ages, gender, and resolution (Mehrabi et al., 2021). Additionally, there has even been a proposal by Google researchers that suggests that "a biased dataset can be perceived as an unbiased dataset which has gone through manipulation by a biased agent" (Horev, 2019). In the proposal, they outline a methodology to re-weight a biased dataset to fit an (theoretical) unbiased dataset, which is characterized by the following metrics: demographic parity, disparate impact, equal opportunity, and equalized odds (Horev, 2019). While a completely unbiased dataset is impossible to achieve given that social context is reflected in it, it is still possible that no one group of people is being marginalized in it. And regardless of modifying a dataset to add more diverse values or propelling it through a mathematical transformation to turn it into a partially unbiased dataset, human intervention is needed at every step of the data preparation process and algorithm training process to provide checks and balances along the way.

The Social Responsibility of IT professionals

The STS SCOT framework emphasizes the importance of human involvment when eliminating bias injection in machine learning models. It prompts questions like "Who are the relevant social actors?", "What are their interests and relative amounts of power?", and "Which people need to approve this algorithm?" (Green & Viljoen). When identifying the relevant social actors in the development of a machine learning algorithm in industry, the main actors involved include a Data Science team manager, developers, business analysts, interactive users, MLOps (machine learning operations) engineers, IT professionals, and compliance professionals (Tamagnini & Winters, 2020). All of these actors have technical backgrounds and work together in a team to meet the business goals of a specific company or person. Unless it is their job or an assigned responsibility to the vet for bias, bias will not be checked. Therefore, they lack incentives to ensure that their work has positive social impacts. In fact, many computer scientists have claimed "I am just an engineer" and "Our job isn't to take political stances", highlighting the lack of social responsibility they feel (Green, 2021). Likewise, in academia, research with societal connotations is invalidated and deemed profitless, which can be seen with the limited internal data science roles present in governments and nonprofit organizations (Green, 2021). Clearly, the lack of incentives for computer science professionals in industry and academia impedes their willingness to tackle projects with societal connotations or vet existing projects for societal connotations.

The first step towards combating this would be to instill interest in computer scientists and other technical professionals to be political actors in developing algorithms, which requires collaboration with external communities (Green, 2021). In other words, it requires the engagement of stakeholders. One social group or stakeholder that needs to be engaged with the most and not be neglected is the minority groups targeted by biased algorithms. The realization that algorithms negatively impact minority groups generally comes from statistical analyses instead of from consulting with a minority who has experienced AI bias, highlighting again how quantification is prioritized over human word. For instance, many women, specifically women of color, have expressed distaste towards the lack of feminist data that algorithms train on and the failure of facial recognition softwares to recognize black females. In fact, feminist artist and game maker A.M. Darke questioned 'Why should a few hundred mostly white, mostly men dictate the procedures that bind us and create or limit our agency in the world?', and created an algorithm that is overtly biased against white men to highlight the problem of AI bias (Healey, 2020). Similarly, poet and academic Joy Buolamwini expressed that she wants a "world where technology works for all of us, not just some of us." In response, she also founded the Algorithmic Justice League which collects people's experiences about bias in AI and uses them to audit software and create more inclusive data sets (Gruber). Computer scientists and IT professionals should do the same and actually listen to those being affected by AI bias to find ways to improve their user experiences.

Additionally, computer scientists should use practical reasoning to hold themselves and their companies accountable to the effects of their developed algorithms. In other words, it should be a collective effort. All algorithm developers should build solidarity towards eliminating racial and gender inequity as outcomes of the algorithms they develop. As an example, in 2019, thousands of computer science students across universities in the US boycotted Palantir as a potential employer due to its partnerships with Immigration and Customs Enforcement (ICE) (Green & Viljoen). Similar behaviors should be emulated by computer science professionals in the workplace. To begin with, developers should increase communication with managers and other technical team members to gain transparency about the larger goals of their algorithms. Then, if they sense that their algorithms have potential for bias, they can either assemble with other team members to force change or construct a best practices guide to determine if an algorithm should continue to be worked on. However, doing so could risk their employment. Alternatively, developers can individually voice their concerns to upper management and provide substitutions or modifications to the algorithms. While none of these tactics guarantee that biased algorithms won't be deployed, they are a step towards normalizing conversations about biased algorithms and increasing accountability of those involved in their creation in the workplace.

Opposition to Algorithmic Realism

Though many computer scientists vouch for a realism approach to debiasing machine learning algorithms, there has been some controversy regarding the introduction of subjectivity in algorithm development. In traditional human-decision making, to harmonize the outcomes of a decision between different groups of people requires those groups of people to be treated differently. This notion is defined as disparate treatment and clashes with the idea of disparate impact, which refers to when certain practices with seemingly neutral results actually disproportionately affect a protected group (Society for Human Resource Management). Generally, with most institutions and human-decision making, the tension between disparate treatment and disparate impact is bypassed with case-by-case workarounds. However, Computer Scientist and associate professor at Princeton Arvind Narayan has voiced that finding creative case-by-case workarounds doesn't scale well for algorithm-based decision making when the sole purpose of machine learning algorithms is to have uniform and automated ways to make decisions (Narayanan). Moreover, human involvement in every step of constructing an algorithm would simply result in greater time and costs. Additionally, many have argued that the involvement of humans would result in significantly more biased outcomes as humans have been recognized as remarkably bad decision makers (Miller, 2019). For instance, research done by psychologists Paul Meehl and Robyn Dawes in the 1950s demonstrated that simple mathematical models outperformed supposed experts in predicting clinical outcomes (Miller, 2019). Consequently, the use of quantifiable metrics to evaluate algorithms has been supported as it has been claimed that using zero or limited statistical/significance tests would result in worse outcomes. Evidently, increased subjectivity does not completely resolve algorithm bias and is more feasible in principle. Perhaps, developers should aim to strike a balance between subjectivity and objectivity in developing algorithms while constantly checking for bias and marginalization in the process.

Conclusion

Machine learning algorithms, specifically those utilized by image and text classification systems, enable racial and gender inequity from the cyclical nature of cognitive bias injection they undergo. This is mainly due to their adherence to algorithmic formalism, which contrastingly emphasizes a level of objectivity when constructing and deploying an algorithm. Furthermore, objectivity is closely related to social power, which big tech executives and people of power take advantage of to fuel their privilege and status when they deploy discriminatory AI technologies. Consequently, the shift from "algorithmic formalism" to "algorithmic realism" is imperative to expanding the bounds of algorithmic reasoning and eliminating biases. Utilizing social context and human checks at every step of the algorithm development process is crucial in preventing biases to seep into algorithm outcomes. As discussed earlier, some realist methods include bias impact statements, the engagement of stakeholders, and the transformation of biased datasets into partially unbiased ones. However, it should be noted that the engagement of stakeholders, specifically with non-technical stakeholders is paramount. The voices and opinions of the minority groups affected, in particular, need to be taken into account in the development process. In doing so, computer scientists have great social responsibility as political actors in sociotechnical systems. Although opposition to algorithmic realism exists, a level of subjectivity and human intervention is needed in the algorithm development process. Of course, they should exist in conjunction with some objective metrics and measures. Overall, computer scientists and tech professionals around the world must build solidarity in making it a priority to eliminate outdated and current gender/racial prejudices present in algorithmic outcomes for good.

References

Beer, D. (2016) Metric power. London: Palgrave Macmillan.

Breland, A. (2017, December 4). *How White Engineers built racist code – and why it's dangerous for black people*. The Guardian. Retrieved May 2, 2022, from https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police

Caliskan, A. (2022, March 9). *Detecting and mitigating bias in Natural Language Processing*. Brookings. Retrieved March 29, 2022, from https://www.brookings.edu/research/detecting-and-mitigating-bias-in-natural-language-processing/

Cambridge Dictionary. (n.d.). *Objectivity*. OBJECTIVITY | definition in the Cambridge English Dictionary. Retrieved May 2, 2022, from https://dictionary.cambridge.org/us/dictionary/english/objectivity

Crawford, K., & Paglen, T. (n.d.). *Excavating AI*. The Politics of Images in Machine Learning Training Sets . Retrieved May 2, 2022, from https://excavating.ai/

Data Justice Lab. (n.d.). *Advancing civic participation in algorithmic ... - data justice lab*. Retrieved May 2, 2022, from https://datajusticelab.org/wp-content/uploads/2021/06/PublicSectorToolkit_english.pdf

Green, B. (2021, September). *Data Science as political action - arxiv.org*. Retrieved March 29, 2022, from https://arxiv.org/pdf/1811.03435.pdf

Green, B., & Viljoen, S. (n.d.). *Algorithmic realism: Expanding the boundaries of ...* Retrieved March 29, 2022, from https://www.benzevgreen.com/wp-content/uploads/2020/01/20-fat-realism.pdf

Gruber, B. (n.d.). *How artists are hacking bias in algorithms*. Goethe Institut. Retrieved May 2, 2022, from https://www.goethe.de/prj/k40/en/kun/aia.html

Hao, K. (2020, April 2). *This is how AI bias really happens-and why it's so hard to fix*. MIT Technology Review. Retrieved March 29, 2022, from https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Healey, B. (2020, December 22). *This artist is creating a "belligerent algorithm" to expose Ai Bias*. Verdict. Retrieved May 2, 2022, from https://www.verdict.co.uk/ai-bias-artist/

Horev, R. (2019, February 9). *Identifying and correcting label bias in Machine Learning*. Medium. Retrieved March 29, 2022, from https://towardsdatascience.com/identifying-and-correcting-label-bias-in-machine-learning-ed177d30349e

Hagendijk, R. (1999). An Agenda for STS: Porter on Trust and Quantification in Science, Politics and Society [Review of *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, by T. M. Porter]. *Social Studies of Science*, *29*(4), 629–637. http://www.jstor.org/stable/285655

Hamilton, A. M. (2021, July 7). *Silicon Valley pretends that algorithmic bias is accidental. it's not.* Slate Magazine. Retrieved May 2, 2022, from https://slate.com/technology/2021/07/silicon-valley-algorithmic-bias-structural-racism.html

Ingold, D., & Soper, S. (2016, April 21). Bloomberg.com. Retrieved May 2, 2022, from https://www.bloomberg.com/graphics/2016-amazon-same-day/

Kliegr, T., Bahník, T., & Fürnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, *295*, 103458. https://doi.org/10.1016/j.artint.2021.103458

Lee, N. T., Resnick, P., & Barton, G. (2022, March 9). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings. Retrieved March 29, 2022, from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

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

Miller, A. P. (2019, November 21). *Want less-biased decisions? use algorithms.* Harvard Business Review. Retrieved March 29, 2022, from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

Narayanan, A. (n.d.). *Tutorial: 21 fairness definitions and their politics - youtube*. Retrieved March 29, 2022, from https://www.youtube.com/watch?v=jIXIuYdnyyk

Raji, I. D., & Fried, G. (2021, February 1). *About face: A survey of facial recognition evaluation*. arXiv.org. Retrieved March 29, 2022, from https://arxiv.org/abs/2102.00813

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018, April). *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. Algorithmic impact assessments - AI now institute. Retrieved May 2, 2022, from https://ainowinstitute.org/aiareport2018.pdf

Samuel, S. (2022, April 19). *Why it's so damn hard to make ai fair and unbiased*. Vox. Retrieved May 2, 2022, from https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence

Schwemmer, C., Knight, C., & Bello-Pardo, E. (n.d.). *Diagnosing Gender Bias in Image Recognition Systems*. Retrieved March 29, 2022, from https://journals.sagepub.com/doi/full/10.1177/2378023120967171

Society for Human Resource Management. (n.d.). What are disparate impact and disparate treatment? Retrieved May 2, 2022, from https://shrm.org/ResourcesAndTools/tools-and-samples/hr-qa/Pages/californiacomplianceourcompanyhas100employeesonly1employeeworksinca.aspx

Stemwedel, J. D. (2013, February 26). *The ideal of objectivity.* Scientific American Blog Network. Retrieved May 2, 2022, from https://blogs.scientificamerican.com/doing-good-science/the-ideal-of-objectivity/

Solly, M. (2019, September 24). *Art project shows racial biases in artificial intelligence system*. Smithsonian.com. Retrieved March 29, 2022, from https://www.smithsonianmag.com/smart-news/art-project-exposed-racial-biases-artificial-intelligence-system-180973207/

Tamagnini, P., & Winters, P. (n.d.). *Data Science in Action: Stakeholder Collaboration - YouTube*. Retrieved March 29, 2022, from https://www.youtube.com/watch?v=yIFgrtTqdAE

Wikipedia contributors. (2021, November 14). *Algorithmic bias*. Wikipedia. https://en.wikipedia.org/wiki/Algorithmic_bias