Decoding how Cognitive Bias Becomes Algorithmic Bias in Artificially Intelligent Decisioning Systems

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Coby Chiu

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

Joshua Earle, Department of Engineering and Society

STS Research Paper

Introduction

In recent decades, the imperfection of the human mind has become a widely studied topic whose findings encourage us to be more cognizant of the way we see and respond to the world (Kahneman, 2011). Researchers in cognitive psychology are interested in understanding why people are prone to decisions or behaviors that are not backed by logic. Others are also interested in discovering what biological advantages, if any, those actions offer (Haselton, 2006). As work is done to figure out the mechanisms of human intelligence, the rise of the technological age spurs research in a related area: artificial intelligence (AI).

Computing devices are generally thought to be free from logical error in all the ways humans are not. One assumes that typing an equation into a calculator will return the correct result. Automating a robot to assemble parts in a factory should produce something more accurately than a person could. Thus, it must be that having a machine make a choice should leave out the faults of a human decision-maker. Unfortunately, this tends not to be the case. A study conducted in the MIT Media Lab found that several commercially used classification algorithms are more error-prone when classifying darker-skinned individuals, especially women (Buolamwini, 2018). When we consider how prevalent facial recognition technology is and is becoming, perfecting that technology must be an industry standard. If so, why is it not our reality?

In this research paper, I will investigate this question: how is it possible that lines of code and numbers in a computer could exhibit the same logical flaws a human does? The goal is to create an understanding of algorithmic bias that draws from both a computer science and cognitive science background.

In the following sections, I will first outline the methods and STS frameworks used to compile information. I will then introduce background information on the topic. Next, I will explore case studies of decisioning algorithms and analyze the algorithmic bias present in them. Finally, I will discuss the implications and future work of this topic.

Methods

I employed finding, reading, and synthesizing previous literature as my method for gathering information. To begin my search, I wanted to develop a good foundation in the history of my topic. I looked for books and papers relating to the keywords of "artificial intelligence" and "cognitive bias" within the University library and scholarly journals. The first primary source used for this topic is Affect and Artificial Intelligence by Elizabeth Wilson. This book investigates the early days of AI and argues that newer developments in computation, namely the entry of cognition and minds, can be traced back to the origins of the science.

To support my understanding of biases in both machines and people, I drew from texts I read in past psychology and STS classes. These readings were supplemented by research papers I could find relating to documented instances of algorithmic biases and studies on AI ethics. Joy Buolamwini's "Gender Shades" served as a model for the type of content I wanted . The paper highlighted shortcomings in the technology and where they detected the origins of the faults. I sought to find other case studies that offered similar insights and other articles to fill in gaps when necessary.

I use the case studies my research uncovered, and I analyze them through the lens of Black/race critique. The case studies are instances of bias in facial recognition software, bias in criminal identification, and bias in applicant tracking systems. Because the issues found in AI

decisioning tend to disproportionately impact people of color, specifically Black people, I found the Black/race critique framework to be especially relevant.

Background on Cognitive Bias

What is Cognitive Bias?

Cognitive biases are subconscious logical flaws that cause a person to respond to a situation in an erroneous manner. They are a result of short cuts, known as heuristics, the brain uses to make decisions more efficiently (Soleimani, 2021).

In his book *Thinking Fast and Slow,* Daniel Kahneman presents two systems of thought: System 1 – the quick, reactionary process, and System 2 – the slow, rational process. He suggests that the first system is responsible for the mind's tendency to seek out heuristics. Its purpose is to expedite decision making, aiming for the most efficient solution to the task at hand. However, this system is also automatic and prone to making mistakes. An example of this system at work is glancing at an object and instantly seeing it is a person even though it is not. One way to bypass System 1 thinking is to slow down; increasing the time of the thought forces the brain into System 2 thinking (Kahneman, 2011). System 2 serves to be deliberate and logical in its performance, allowing one to carefully navigate a task.

There are many theories as to why humans exhibit this way of thinking. Anthropologists frame the reactionary qualities of cognition because of evolution; ancestors to humans likely relied on quick decisions for survival (Ellis, 2018). Psychology points to learning as the culprit, where a person might acquire a certain way of thinking and behaving due to past experiences (Ellis). Deriving an explanation for the way people think is the driving force of an overarching field called cognitive science.

The Study of Cognitive Science

The desire to understand the functions of the human mind dates to, at minimum, ancient Greece (Thagard, 2023). Philosophy, psychology, anthropology, and similar fields studied by people worldwide have branched from that same desire to study what we now call cognition. Centuries later, psychologists in the 1950s began to liken the function of human memory to computational functions (Hogan, 2012). This way of thinking paralleled the introduction of the term "artificial intelligence," and eventually resulted in a new field of study: cognitive science.

Cognitive science is an interdisciplinary field that studies cognition while drawing from the disciplines of cognitive psychology, philosophy, neuroscience, computer science (specifically AI), linguistics, and anthropology (Eckardt, 1993). As a result, cognitive science is highly relevant to discussions on AI and biases.

Background on Artificial Intelligence and Machine Learning

Where it Started

Alan Turing was a British mathematician who is known as the "Father of Modern Computer Science" for his vision of and contributions towards the field of computer science as we know it today (Wilson, 2010). He worked on a team at the National Physics Laboratory whose goal was to build an automatic computing machine. After a brief period, he left the project citing that he wanted to focus more on creating a machine that would model the human brain rather than a computer with practical applications. This sentiment carried through all of Turing's work as a computer scientist. He began asking more philosophical questions about intelligence and cognition, one of his most salient being, "Can machines think?" (Turing, 1950).

The term "artificial intelligence" was first used in the 1950s to describe machines that operated above a rudimentary level (UNESCO, 2019). AI is defined as "the science and engineering of making intelligent systems" (Kersting, 2018). Since its conception, the field has

expanded and divided into different subsets, one being machine learning (ML). ML is a method by which to get a machine to improve its performance on a given task. Both AI and ML aim to develop intelligent computers that are designed to process data in a way that accomplishes a semantic, or meaningful, goal (Kersting, 2018). That goal might be to label individual objects in an image or select the most efficient path to a destination.

How Do These Technologies Work?

A machine learning algorithm takes sets of data as input, extracts semantics from that data, then produces a model based on its learning. These models are comprised of rules, procedures, and functions that can be used in a general application to complete a decisioning task (Kersting, 2018). Training such an algorithm is typically accomplished through one of several learning techniques such as supervised or unsupervised learning. The former presents the machine with inputs that are already labeled with information while the latter requires the machine to extrapolate patterns from unlabeled data.

ML is a process that is used in artificial intelligence to facilitate learning. One approach to AI, called “neural networks” or “deep learning,” is functionally similar to the way a human brain works (Hardesty, 2017). It accomplishes machine learning through a series of deeply connected points that process information organized into different layers. The training data is received by an input layer before being propagated to successive layers before being output as something semantically significant.

In the decisioning technologies I will discuss in the next couple sections, algorithms rely heavily on machine learning to prime a system to produce a decision about its environment. In each instance, a cycle of obtaining data, feeding it to software, and receiving a conclusion is expressed.

Algorithmic Bias and Job Applications

Background

Particularly in positions that net large numbers of applications, employers turn to automated means of filtering out qualified candidates (Raghavan, 2020). In fact, almost 75% of recruiters or hiring committees rely on applicant tracking software (ATS) (Deshpande, 2020), and consequently, only about 28% of resumes ever make it to an actual person (Wilson, 2018). ATS works by training an algorithm on a dataset that includes labels on what content is and is not favorable to appear on a resume (Soleimani, 2021). Its goal is to screen the written content of each resume presented to it and determine if the resume should be passed onto the next round. The guidelines for what constitute an ideal candidate are established by hiring managers and formalized by the developers of the software (Soleimani). Often, the software also checks for language that matches that of the job description.

Studies conducted before ATSs became a standard practice found that resumes with white-passing names received 50% more callbacks than names perceived as Black (Wilson). While ATSs might not take the names on resumes into consideration the same way a human might, the results of the algorithm still exhibit bias based on gender and ethnicity (Wilson, 2018; Deshpande, 2020).

ATS Datasets

ATSs are frequently used for applicant filtering due to their efficiency and perceived objectivity (Wilson, 2018). However, they have proven to still favor certain demographics more than others (Soleimani, 2018; Deshpande, 2020). The reason for this is clear when we note that ATSs are trained using datasets of previous decisions made by managers.

Hiring managers tend to be biased in favor of white applicants (Wilson, 2018). This display of prejudice is a demonstration of that person's cognitive biases creating a perceived space that differs from reality (Rastogi et al., 2022). When a decisioning system is trained using a precedent set by a human's subconscious choices, the machine learns that perceived space to be a ground truth. The result is software that is programmed to make the same logical errors the people before it did (Soleimani, 2018).

A different approach to training these systems involves giving developers a set of criteria for the ideal candidate. Developers are then tasked with formalizing the dataset and producing algorithms based on their guidelines. In this scenario, bias might be introduced in either the managers' or developers' biases in either formulating the algorithm or preparing the data (Soleimani, 2018). Human resources managers who provide the baselines for a good candidate are at risk of including their personal preferences. They might also lack firsthand knowledge of the skills required for the position, skewing the model. Developers who handle the formalization of the algorithms might make assumptions to fill in the gaps of their datasets. They then would introduce their own set of cognitive biases to the model (Soleimani).

Algorithmic Bias and Law Enforcement

Background

Policing in recent years has moved away from purely reacting to phone calls reporting crimes, instead favoring proactive surveillance of areas deemed to be higher risk (Brayne, 2017). The decision on which neighborhoods should be considered "hot spots" for crime tends to be made by a computer algorithm in many police departments (Wilson, 2018). These algorithms differ on an individual level, but generally, they operate by logging data on individuals and their interactions with the police (Brayne, 2017).

The Los Angeles Police Department (LAPD) uses a software called PredPol which runs on a model that assumes crime is more likely to happen in the areas around the location of a prior crime (Brayne, 2017). As crimes are logged within the department, the algorithm uses data about the incident and produces a prediction of where future crimes might take place. These take the form of 500 by 500 square foot boxes drawn onto maps that are then handed to officers before they go on patrol (Brayne).

Predictive Policing Datasets

The dataset considered by predictive policing software is the crimes and surface level details of those crimes in a certain area. The issue is that to PredPol and other similar proprietary software, "crime" can mean anything from jaywalking to murder. There is an option to filter out the lesser crimes from the prediction model, but most departments opt to include all data due to the belief that more is better (Wilson, 2018). When more data points on smaller crimes are accounted for, the police presence in areas where shoplifting or parking violations occur increases. When police presence increases, the likelihood of someone being caught doing something illegal increases, even if doing the same act in a less surveilled area would go unnoticed (Wilson). This incident of crime gets added to the system, and the cycle becomes a feedback loop. Thus, the same neighborhoods are kept under watch, and the same people are trapped into run-ins with the police. These people tend to disproportionately be Black, Latino, houseless, or impoverished (Wilson).

Algorithmic Bias and Facial Recognition

Background

A study conducted by the MIT Media Lab compared the classification abilities of commercial facial recognition software built by three different companies: Face++, Microsoft,

and IBM. The software was to label images of people's faces and determine their gender and whether they had darker or lighter skin. The researchers found that the software was most accurately able to classify people with lighter skin and people who were men. Speaking in terms of intersectionality, lighter men were correctly classified the most while darker women were the most likely to be misclassified (Buolamwini, 2018). The accuracy of the results ranges from 87.9% to 93.7%, but the classification is near perfect for light-skinned men while it misclassified about 20.8% of dark-skinned women (Buolamwini).

The findings of this study are replicated in others. A series of reports by the National Institute of Standards and Technology tested the accuracy of certain facial recognition software and found that people of color and women were most likely to be misidentified (Grother, 2019).

Case Study: Robert Williams

A Black man was caught on a security camera stealing thousands of dollars' worth of expensive watches. At the beginning of 2020, a different Black man was arrested for the crime. Robert Williams became a victim of misidentification by a facial recognition software used to find the identity of the man who stole the watches (Perkowitz, 2021). He was kept overnight in a detention center following his arrest, his fingerprints and mugshots taken, and was interrogated the next day. After the detectives finally compared Williams to the man in the video, they realized they had the wrong man, but it took them thirty hours and Williams posting a $1,000 bond to see their mistake (Hill, 2020).

Faults in Training Images

Since the invention of the camera, those with lighter skin have been put at an advantage in the world of photographic representation. Lenses have been designed to capture lighter colors

more effectively, and we see this to be an issue in an age where those lenses are used to detect identities (Leslie, 2020).

Facial detection and recognition technologies (FDRTs) rely on datasets of thousands of images to learn how to classify people. These images are tagged with metadata called "labels" that describe aspects of the person in the image; for example, "white" or "woman." Variations in the clarity of these images and the specificity of the labels assigned lead to inconsistencies in the accuracy of FDRTs (Leslie). As a result, false positives, or the incorrect association of two subjects, are highest in African and East Asian people and lowest in Eastern Europeans (Grother et. al., 2019). Technology that tends to conflate two people of similar visible ethnicities highlights why people like Robert Williams are wrongly convicted of crimes they had no affiliation with.

In addition, the diversity makeup of standard datasets is not always equal (Han and Jain, 2014). If a team is not able to pick apart and optimize a given dataset, they risk using thousands of images that feature more white people and men over everyone else. Training a model on such data optimizes its use for a specific type of person, resulting in the metrics that favor white men (Buolamwini, 2018).

The Impact of Biased Training Data

The bias in algorithms used in ML or AI like the ones explored in the prior case studies can in part be linked to the data used to train those models. Training in ML requires a lot of data to create a good model, and there are publicly accessible datasets built to provide those resources. These sets are curated for specific use cases and can be anything from statistics to images of people or objects. Other models rely on using data scraped from the internet or

databases. While these data are necessary for machine learning, they are also prone to introducing bias to the system.

In the case of applicant tracking systems, the technology learns from past decisions made by hiring managers. Predictive policing bases its suggested surveillance map on prior interactions between people and officers. Facial recognition tasks learn from a set of images whose diversity and labeling are in the hands of humans. These technologies have demonstrated flaws in their means of gathering information. As a result, the decisions they make should not be trusted to be completely objective. However, we note that a dependency on the results generated by AI is also a contributing factor to the bias in decision-making.

Bias in Decisions Made With AI

Humans tend to rely too heavily on the word of technology as objective. This phenomenon is known as anchoring bias, where a person is inclined to stick to the perception provided by an anchor (the AI decision) and does not explore other possibilities (Rastogi et al., 2022). Automation bias is a related occurrence which describes a person's blind trust in AI. Both automation and anchoring bias highlight that humans have a predisposition to believing what a machine says is true and trusting the decision wholly (Rastogi et al.). These biases are problematic when coupled with the fact that AI is not always correct or objective.

It had been visually clear that Robert Williams was not the man in the shoplifting video, yet he was still taken in and interrogated for more than a day. This incident and others like it are direct results of law enforcement blindly trusting AI without verifying that the decisions are correct. From this case study, we find that bias exists not only in the AI decisions, but also in what people decide to do with those decisions.

Ethical Implications of Algorithmic Bias

In many ways, the imperfections of decisioning AI mirror the prejudices found in human society. Unfortunately, these imperfections have real life consequences in the form of exacerbated racism and sexism (Noble, 2018). Robert Williams is not the only person who has been wrongly convicted of a crime following an incorrect AI decision (Leslie, 2020). Stereotypes and biases presented in society are greatly reinforced by the AI that learned from them (Noble, 2018). The more we introduce AI into all facets of life, the more likely that the biases it presents will have grave consequences. In systems such as law enforcement that are already scrutinized for their histories of injustice, discriminatory technology is the last thing marginalized folks need.

The United Nation Educational, Scientific, and Cultural Organization (UNESCO) released a study on the ethics of AI. In it, they offer concerns regarding the future of education, public policy, and society in general. The report notes the amplification of biases through AI systems. One example they provide is the use of female voices in AI assistants which might emphasize societal stereotypes towards women being subservient (2019). They also note that training data is also subject to reflecting societal biases. UNESCO does, however, also point out that there is a lot of potential for good in the rise of AI. It is still recommended that a code of ethics is applied and adhered to combat biases during development.

Proposed Solutions

Dismantling the notion that AI is inherently objective is the first suggestion to minimize the impacts of algorithmic bias. Kahneman's work on cognitive bias posits that increasing the time it takes to decide forces the brain to slow down and rely on more rational thought processes. Encouraging this same behavior even when a machine offers its input would reduce the effects of anchoring and automation bias (Rastogi et al., 2022).

Developers must also be mindful of the data that is used to train their models. Often, the most accessible databases are also ungeneralizable, so it is important to make necessary changes to the scope of the training content. Namely, training that involves identifying human demographics should train equally on different groups such as ethnicity or gender.

Future Work

Extensions of this research could dive deeper into feasible solutions for the problems presented by cognitive and algorithmic bias. This paper sought to create an understanding of how these biases occur more than the pressing concern as to how to fix them. Another possibility would be to explore more areas of AI and ML and how bias is presented in them. The goal would be to confirm the findings of this paper and offer a holistic general explanation for algorithmic bias.

AI is a rapidly evolving field whose research implications multiply constantly. As artificial systems become more and more intelligent, the need for interdisciplinary approaches to studying them increases. An interesting vein of research might explore what, if any, biases have formed that are unique to AI. This work would draw from Turing's original notion of AI as a model of the human brain to see if developers have or could create systems that might develop their own biases. Such a discovery would push the boundary of what is considered sentient.

References

Brayne, S. (2017). Big Data Surveillance: The case of policing. *American Sociological Review*, *82*(5), 977–1008. https://doi.org/10.1177/0003122417725865

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Retrieved October 12, 2022, from https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Eckardt, B. V. (1993). *What is cognitive science?* MIT Press.

Ellis, G. (2018). So, what are cognitive biases? *Cognitive Biases in Visualizations*, 1–10. https://doi.org/10.1007/978-3-319-95831-6_1

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face recognition vendor test part 3: https://doi.org/10.6028/nist.ir.8280

Han, H., & Jain, A. (2014). (rep.). *Age, Gender and Race Estimation from Unconstrained Face Images* (pp. 1–2). East Lansing, MI: Dept. Computer Science and Engineering.

Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News. Retrieved October 12, 2022, from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review*, *10*(1), 47–66. https://doi.org/10.1207/s15327957pspr1001_3

Hill, K. (2020, June 24). Wrongfully Accused by an Algorithm. *The New York Times*. Retrieved March 14, 2023, from https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hogan, P. C. (2012). *Cognitive science, literature, and the arts a guide for Humanists*. N.Y.

Kahneman, D. (2011). *Thinking Fast and Slow*. Farrar, Straus and Giroux

Kersting, K. (2018). Machine Learning and Artificial Intelligence: Two fellow travelers on the quest for intelligent behavior in machines. *Frontiers in Big Data*, *1*. https://doi.org/10.3389/fdata.2018.00006

Kinsey, M. J., Gwynne, S. M., Kuligowski, E. D., & Kinateder, M. (2018). Cognitive biases within decision making during fire evacuations. *Fire Technology*, *55*(2), 465–485. https://doi.org/10.1007/s10694-018-0708-0

Korteling, J. E., Brouwer, A.-M., & Toet, A. (2018). A neural network framework for cognitive bias. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01561

Noble, S. (2018). *Algorithms of Oppression.* NYU Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Preliminary Study on the Ethics of Artificial Intelligence. unesco.org. (2019). Retrieved March 14, 2023, from https://unesdoc.unesco.org/ark:/48223/pf0000367823

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–22. https://doi.org/10.1145/3512930

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372828

Soleimani, Melika & Intezari, Ali & Taskin, Nazim & Pauleen, David. (2021). Cognitive biases in developing biased Artificial Intelligence recruitment system.

Thagard, P. (2023, January 31). *Cognitive science*. Stanford Encyclopedia of Philosophy.

Retrieved April 8, 2023, from https://plato.stanford.edu/entries/cognitive-science/

Turing, A.M. (2009). Computing Machinery and Intelligence. In: Epstein, R., Roberts, G., Beber, G. (eds) Parsing the Turing Test. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6710-5_3

Verbeek, P. & Parizeau, M. (2019). Preliminary Study on the Ethics of Artificial Intelligence. Retrieved October 12, 2022, from https://unesdoc.unesco.org/ark:/48223/pf0000367823

Wilson, E. (2010). Affect and Artificial Intelligence. University of Washington Press Seattle, WA