The Human Face of Algorithms
how human biases translate into technology

Lily Chu
Wellesley College
106 Central Street, Wellesley, MA, 02481, USA

ABSTRACT
The relationship between society and technology is dynamic and largely shaped by the programmers and innovators who create the technology we use. With these rapid advances come problems such as algorithm biases that have yet to be addressed by programmers or policy makers, let alone understood by the general public. Algorithm biases have larger implications than realized. Examining these algorithms reveals that the natural languages used by those who write them and the web they explore carry their own biases and thus drastically impact the way the technology we're creating and training process information in this increasingly digital world. Perhaps most alarmingly, the extent to which the dangers of these inherent and exacerbated biases are understood by the general public is small, furthering the problems that arise from the already limited participation in the shaping of this socio-techno landscape as technology moves forward. This paper presents potential solutions to algorithm biases and how we hope this will positively change how the larger public interacts with technology.

KEYWORDS
Bias, Prejudice, Algorithm Bias, Artificial Intelligence

1. INTRODUCTION

Technology driven by algorithms and artificial intelligence (AI) is becoming increasingly present in our lives, living with access to as much information in a day as that of a 15th century lifetime. With more data and the ability to process it through algorithms, we are able to solve larger problems than ever before. And yet, we are also creating larger problems than ever before, forgetting that algorithms are only as intelligent as the humans that created them. In this way, in our enthusiasm for greater efficiency and efficacy comes the enshrinement of some of the best and worst of humanity, and as with biases it is often the latter that becomes the source of our problems If biases inside algorithms that make important decisions go unnoticed, there could be serious, negative consequences. Algorithms can be used to determine anything from the order of content in your Facebook feed to who gets hired for certain jobs or even who spends how much time in prison, spiraling out of personal spheres into public domains as we question the value of tax money and efficacy of our justice system. But even beneath this is a more foundational problem. The problem is most Americans are unfamiliar with how present and influential algorithms are in their lives. 57% of Americans say they have never heard of anything like computer generated algorithms to analyze job candidates. However, when Americans do find out that computer generated algorithms have such a significant impact on their lives like deciding whether they receive a certain job or not, over 77% of respondents expressed that they are not enthusiastic about the prospect, 67% said they are very worried about algorithms making these decisions, and 76% said they would not want to apply to a job if they knew that the decisions were being made by a computer (Smith et al., 2017). With these algorithms making more and more decisions in our everyday lives, it is important to address the ways biases and worse, prejudices, are influencing our world and how we can fight this increasingly

pertinent problem. Algorithms have the power to connect the world, empower individuals, solve larger problems, and improve quality of life. But failing to address the negative fallout and not correcting errors changes the question from how much life can be improved by technology to how much life there is left to live as we relinquish the power to make decisions to algorithms and the biases that come with them.

2. ALGORITHM BIASES

2.1 Types of Algorithm Biases

There are three different types of biases that arise from algorithms. The first type is pre-existing bias, which has roots in social institutions, practices, and attitudes. Pre-existing bias can enter a system either through the explicit and conscious efforts of individuals who played a large part in the creation of the algorithm or society at large. The second type is technical bias, which arises from technical constraints or technical considerations. There can be limitations in computer technology including hardware, software, and peripherals, algorithms that fail to treat all groups fairly under all significant conditions, imperfections in pseudorandom number generation or in the misuse of pseudorandom numbers, or attempts to make human constructs such as discourse, judgments, or intuitions amenable to computers. The final type of bias that can arise is emergent bias, which happens in a context of use with real users. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. If there is an emergence of new knowledge in society that cannot be or is not incorporated into the system design, or if the population using the system differs on some significant dimension from the population assumed as users in the design, this bias can occur (Friedman 1996).

2.2 Problems that Algorithm Biases Present

The problems that algorithm biases cause can range from discriminatory results on social media feeds, to factors that impact people's everyday lives. It is important to note that algorithm biases and its resulting problems all can have positive or negative impacts, whether they are intentional or not. Although these problems can be quite diverse in effect, they all originate from human prejudices due to the fact that algorithms are created by people. In other words, the problems that algorithm biases present ultimately reflect the attitudes and behaviors of the algorithm's creator(s); however, algorithm biases and its problems cannot be resolved or tracked down by simply taking a look at the behaviors of the creators themselves. As machines are getting more complex, they are also getting increasingly skilled at developing language skills comparable to those of humans. Machines, however, absorb not only these language abilities, but also the deeply rooted biases hidden within patterns of language.

The acquisition of these biases in machines can lead to further problems, such as spreading fake news, making conspiracy theories more available on the Internet, promoting unfair decisions, etc. When it comes to news on the Internet, users rely heavily on search results from Google. Although Google's search result algorithm was not intentionally built to display false news, its algorithm can unintentionally display incorrect articles because of users' tendencies to lie on the Internet. For instance, following the 2016 presidential election, one of the top results on Google regarding the election results was a link to a fake news website that claimed that Donald Trump won the popular vote. This unintentional promotion of unreliable sources can also be seen when one asks Google about a conspiracy theory. Instead of displaying results that refute conspiracy theories, Google shows results that reinforce them, thus

demonstrating that Google's algorithm is vulnerable to manipulation. These problems that algorithm biases present, however, are not limited to news that appears on search results or social media feeds. Everyday lives can be directly affected by problems from algorithm biases; for example, a software that the government uses to measure the probability of repeated offenses can easily put the wrong person in jail. This software, called the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), calculates the likelihood of a convicted criminal re-offending and assigns a correlating "risk score" that judges take into account during sentencing. Research from a year-long investigation by a team of ProPublica reporters revealed that the algorithm for this software only had a 60% accuracy rate. They also discovered that the algorithm was "twice as more likely to call black offenders high-risk when they weren't" (Fortis 2017). The algorithm also did not take external factors into account, such as "disparate rates of policing in minority communities, leading to the biased outcomes" (Angwin et al., 2016). In other words, people can be falsely convicted and put into jail due to the mistakes of an algorithm. These numerous flaws within algorithms ultimately are significant enough to impact not only people's ideas and beliefs, but also people's everyday lives.

2.3 Bias and Prejudice in Natural Language

In language, it is often biases that give us meaning. As with the words "insects" and "flowers", the biases or differences we develop in language are a part of the development of our own understanding of words, and it is in this subjective landscape that we think, speak, and act. However, the introduction of prejudice, defined in Narayanan's paper, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," as "a special case of bias identifiable only by its negative consequences" (Narayanan et al., 2017), is where biases in language become problematic. Let alone algorithms, these prejudices are problematic for humans. The Implicit Association Test documents the existence of this bias which Narayanan and teammates replicated and demonstrated through word embeddings (Greenwald et al., 1998). By representing "the textual context in which a word is found as a vector in a high-dimensional space," Narayanan was able to correlate distances between vectors and semantics or the meaning of words such so that "the associations revealed by relative nearness scores between categories match human biases and stereotypes strongly" (Narayanan et al., 2017). The program GloVe, developed by Pennington et al. (2014), allows the addition of real-world data by Nosek et al. (2002) to be synthesized. The linking of gender-occupation biases and statistical occupational gender gaps embedded in a 300-dimensional vector space shows the words in their true algebraic relationships. The results of these studies and computations are evident in both a chart (not pictured) and graphs through which the determined biases can be visualized. Its correlations, ranging from that of musical instruments to weapons to gender and occupation have varying and far-reaching effects that have not only been proven to exist in the world, but also will be shown to correspond to real-life data and damage through algorithms.

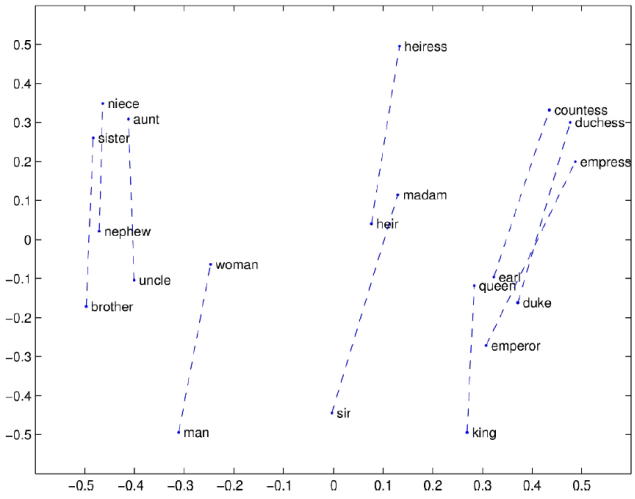

Figure 1. Word relationships are displayed by algebraic lines. These lines demonstrate gender biases as well as larger semantic equity between words.

2.4 Natural Language in Algorithms

It is important, essential even, to remember that this bias exists beyond this study, and as proved by author Cathy O'Neil in her *Weapons of Math Destruction*, prejudice has produced the improbable and absurd statistics of 13% of the American population that is black accounting for 40% of prison populations to be reality. These obvious injustices to undeserving populations points to several things. First, the problems within the algorithms. In many cases, the humans creating the algorithms might not have access to directly related data thus use proxies in place of true data, resulting in the algorithms committing logical fallacies of false cause or false dilemma, conflating correlation with causation or mistakenly limiting the potential solutions. Then, the value placed on different pieces of evidence or data, already flawed in themselves, is determined by a human to create a formula for judging for instance the likelihood of reincarceration. Without transparency and knowing how much value is placed on what bit of information, simple questions such as "Where do you live?" can be maliciously used to determine a person's socioeconomic status and potential for violence. Additionally this human, the coder, has an imperfect sense of morality and hardly one that should surpass that of higher judgment but that still goes unaccounted for simply because we view algorithms as pure mathematics. In this way, we fail to see the underlying prejudice in these flawed formulas, giving them the omnipotence to, as with a 1997 court case against Duane Buck, predict or "mathematically prove" the probability of reincarceration. In this case, Buck was sent to prison for a longer sentence because of "data" and the calculations of an algorithm which determined that black males have a higher rate of reincarceration, justifying its own results by sending Buck to prison in a pernicious feedback loop that ensnares the less powerful into increasing and cyclical loss and poverty and the more powerful into more wealth and power (O'Neil, 2016). This is true not only of racial prejudice but of the likes of the housing bubble burst and consequent recession in 2008 or the increased pressure on standardized testing and fitting into a mold for collegiate acceptance.

Thus, algorithm biases become not only an issue of humans coding their inherent flaws and prejudices into formulas, but the extent to which these algorithms and their biases along with them are

respected by society and those in power, self-justifying in themselves, and powerful over those without power. The humans coding the algorithms are unaware of their own biases and program these into formulas. These algorithmic formulas are regarded as pure mathematics and are sold to those in high positions who seem to neither understand nor care to understand the structure of the assessment, the calculations. In this way transparency and accountability is not only absent but also undesired. These prejudiced algorithms then determine people's job candidacy, potential hazard to society, etc., and it is only through studying the data left behind by the real impacts that we can see the real threat to our justice, democracy, and futures– the prejudiced biases of algorithms.

3. POTENTIAL SOLUTIONS

There is a clear need to form solutions to the biases that algorithms present. As more companies and institutions turn to algorithms to increase efficiency, they must work to counter the rise of biases in them. The United States Association for Computing Machinery Public Policy Council asserts that policymakers should expect institutions that use algorithms to produce a quality of work that is at the same standard as institutions where humans continue to make decisions. The ACM Public Policy Council outlines seven key principles to which algorithm developers should adhere: awareness, access and redress, accountability, explanation, data provenance, auditability, and validation and testing (ACM US Public Policy Council 2017). These standards can be met through the implementation of stricter algorithm development planning procedures, independent review boards, and critical reflections on the efficacy and neutrality of algorithms. Existing projects and movements have already begun to carry out these solutions.

3.1 Preventative Measures

Algorithm developers should anticipate potential biases that may arise in the creation of their algorithm. Clear goals and impacts of unintentional biases should be established. This can be achieved by the implementation of algorithm impact statements which would be similar to the environmental impact statements that large-scale construction projects must produce in order to proceed with their plans. These statements would outline the goals, stakeholders, and expected outputs of a new algorithm (Schneiderman 2016). Such a procedure would compel companies to have the foresight to solve potential problems before they arise as they build their algorithm.

In addition to proactivity from the institution itself, independent review boards should be created to consistently and strictly monitor the functionality and impartiality of algorithms. Similar to how the United States Food and Drug Administration works, algorithm review boards would inspect the effectiveness and quality of algorithms (Schneiderman 2016). The review boards would be composed of knowledgeable individuals–potentially faculty from academic institutions–who have the background necessary to detect algorithm biases (Algorithmic Accountability 2017). To achieve this, they would audit "input data, decision factors, and output decisions and involve documentation" (Algorithmic Accountability 2017). The review board must also have enough authority to investigate incidents of bias, further encouraging companies to avoid careless inclusion of biases into their algorithms (Schneiderman 2016). It is equally important to ensure the impartiality of the review board: "Because inspectors eventually become too close to the system maintainers, regular rotation of inspectors is helpful to ensure continuing independence" (Schneiderman 2016). A third-party regulator would better identify problems of bias in algorithms. Companies should reflect on biases that will inevitably arise and strive to learn from prior mistakes. In order to prevent repetition of the same mistakes and to have references for future projects, companies should carefully track past mistakes and their solutions (Schneiderman 2016). Figure 2 visualizes the cyclical process of review institutions should take in order to minimize algorithm bias.

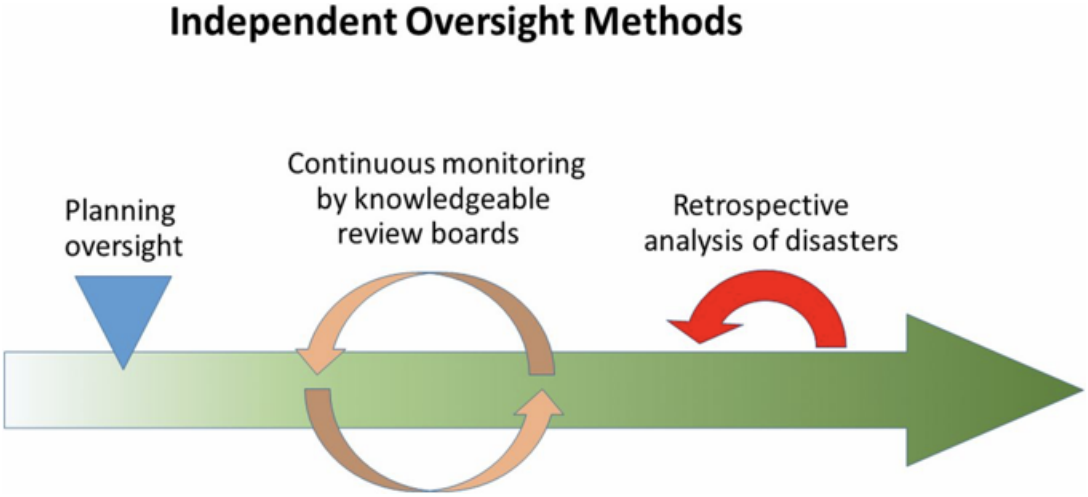

Figure 2. A visual representation of the review process for the detection of algorithm bias (Schneiderman 2016)

Perhaps the most important step humans can take to minimize algorithm biases is education on all levels of algorithm creation; developers and users alike should learn about the sources of algorithm biases and how they can impact a user's experience with technology. The goal of algorithmic literacy in society would be to "enable more individuals to impact information flows and perceive when or if they or others are being marginalized" (Algorithmic Accountability 2017). Unfortunately, these efforts are limited by inequities in access to technology across socioeconomic brackets and the specificity of algorithms in different industries. Furthermore, educational programs on algorithm literacy would have to change as often as algorithms and their associated technologies do; with the dynamic nature of computer science, lessons would be difficult to adapt to the rapidly and ever-expanding field. However, these efforts can and should be realized. One step could be to encourage academic institutions to include ethics courses in their respective computer science degree programs (Algorithmic Accountability 2017). With more awareness of what algorithm biases are and their societal and ethical implications, companies would be able to build less biased algorithms.

3.2 Ongoing Movements

Groups have already recognized the need for decreasing biases in algorithms and have started movements that aim to educate the public about this issue. The Algorithmic Justice League (AJL) promotes awareness among the general public using art and media. They advocate for more algorithm transparency, accountability, and the use of comprehensive and inclusive data sets for the algorithm to use. The AJL employs advocates from all levels of society that interact with algorithms: the developers, academics, and users. They put a strong emphasis on educating the public so that they may also be vigilant and hold companies accountable if they discover a bias in some aspect of their algorithm (Algorithm Justice League 2017).

On a larger scale, the European Union has begun to tackle the issue through the passing of the General Data Protection Regulation (GDPR) in April of 2016. The GDPR is the first piece of legislation to explicitly address the issue of algorithm bias. Its two main goals are "data sanitization and algorithm transparency" (Goodman 2016). Data sanitization involves omitting certain categories from datasets that

could potentially result in the formation of biases. Some of these categories include the race and gender of an individual. The second main goal of the GDPR is to enhance algorithm transparency by granting private users the right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences" from choices made by algorithms (Goodman 2016). The GDPR also requires that developers consider the ethical implications of their work through a process similar to that of the aforementioned algorithm impact statements. It mandates "data controllers to evaluate 'the risks of varying likelihood and severity for the rights and freedoms of natural persons' posed by data processing and to 'implement appropriate technical and organizational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation'" (Goodman 2016). This policy compels developers to seriously consider the impact that their data can have on the people who use their technology and holds them accountable for any discriminatory complications that may arise from algorithms.

Despite the GDPR's good intentions, it has shortcomings. Perhaps its biggest flaw is that there is no clear actor in the execution of these oversights: "The GDPR does not make clear whether monitoring codes of conduct or certification should be conducted by a public, non-governmental or for-profit entity" (Goodman 2016). Due to the lack of knowledge about algorithms among the general public, it is possible that these biases will go unnoticed if it is left up to the public to detect them. For-profit auditing companies present further complications in that, much like the financial sector, "quasi-monopolies" could be created, increasing the costs of auditing. Questions also arise about the unethical relationships that may form between auditors and the companies and institutions they are auditing (Goodman, 2016). Despite the faults with the GDPR, it is still a step in the right direction and should serve as an example for other countries' governments to implement and improve upon.

4. CONCLUSION

This paper has elaborated upon the problems that algorithm biases present, how these problems arise, and what actions can be taken to mitigate these issues. As technology continues to advance, developers should keep in mind the ways in which their new software can work to decrease or exacerbate the already inequitable use of technology in society. The myriad problems that algorithm biases present demand more attention from institutions and the larger public already, and if awareness and lack of accountability continue to decrease exponentially in relation to the technological boom, it is entirely possible that we may wake up one day and realize we are powerless. Though power is not always central, in a functioning democracy and republic, it is undoubtedly essential. From the visionary checks-and-balances of our Founding Fathers to the availability of information at our fingertips, our democracy requires accountability and continued education. As divides deepen, it is of utmost importance to know our democracy and ourselves as there can simply not be one without the other. Algorithm biases are testing this right here, right now. If democracy is to exist in the Digital Age, we must be the ones to know it, advocate for it, and defend it-- something we are failing to do today, something we must do for tomorrow.