Do Advances in Technology Now Amplify Society's Faults? The Many Examples of and Possible Solutions to Artificial Intelligence Perpetuating Bias

Jennifer Good

University of Texas at Dallas

Over the past few decades technology has advanced at an exponential rate. Punch card computers were invented after World War II. Personal computers were developed in the 1980's. They decreased in size to fit into the palm of our hand and became smartphones in the 2000's. Now in the 2010's, we have focused on the computational algorithms in an effort to have computers think and make decisions for us in complex situations such as facial recognition, employment, medicine, and even crime and justice systems. However, in this rush of improvement and advancement, have we strayed from developing a more equitable society? I will explain how across many disciplines machine learning programs are being introduced as the solution to nearly all of society's problems. However, exceedingly many are exacerbating current biases such as racism in our criminal justice system and sexism in job opportunities. The foundation of machine learning programs is data, and though data is abundant in this day and age, without corrections or checks it may reflect current and systemic biases in society.

What is machine learning?

Machine learning is a subcategory of artificial intelligence, but in recent times they have become basically synonymous (Meserole, 2018). Artificial intelligence is a field of computer science in which programmers write code that exhibit intelligence. Early artificial intelligence attempts often used a brute force approach, requiring the creator to include all information and rules for every nuance of decision making. Many important recent advances in the field of AI can be attributed to advances in machine learning. Machine learning relies on statistical analysis and requires a lot of data, so although it was first introduced in the 1950s, its momentum has developed recently with improvements in storage and processing power. Also, with the rise of internet and electronic systems, datasets have accumulated in unprecedented volumes and from

increasingly diverse sources, such as google images, medical records, and police reports to list a few.

Neural networks are an important class of machine learning algorithms (Meserole, 2018). It is so named because to calculate the overall resultant probability, the program breaks down the problem into many true or false nodes that are analogous to neurons in the brain. Each node is connected to others forming a large network, and these connections are usually weighted based on their overall importance in relation to getting the correct result. This means the program must first train on a data set with known expected results. For each entity the weights of connections will adjust based on what probabilities were most related to the correct result. There is a subset of neural networks known as deep neural networks and are more common for complex algorithms such as for facial or object recognition. Deep neural networks differ in that they have layers of nodes between the initial factors and the end results. These internal nodes are created and determined by the program from the first layer, such as in the case of facial recognition where the first layer is specific lines or curves and then the next layer is a combination of those for slightly more advanced shapes like corners or circles. That pattern repeats through many layers until it is left with the final decision about the face as a whole. As will be explained throughout this review, the datasets that algorithms are trained on are very important to the result since each item of the dataset will have an effect on all of the weights. If the training datasets do not adequately reflect statistical regularities in the environment, this can lead to a biased result.

What is bias and in what ways does it occur in machine learning?

Bias is the unequal treatment of a person or group because of perceived characteristics, whether existing or not (Howard, 2017). It can arise most commonly in regards to gender, race,

age, or sexual orientation. Bias may be individual for particular judgements or on a societal scale that can be traced in data such as the demographics of arrest records. The data reflects society and it comes from biases we know we have (explicit) or subconscious ones we may not be aware of at all (implicit). These implicit biases are troubling because they can lead to covertly detrimental effects. In the case of a review from Hall et al. (2015) that analyzed 15 health-related studies, implicit bias was significantly related to patient interactions, treatment decisions, and the patients' health outcomes. And then all of those biases are embedded in the data that may be used to train machine learning algorithms.

In this review I will focus on biases in gender and race because they are the most common and troubling biases that occur in machine learning, and potentially the most resistant to change. Machine learning is typically used because of a perceived fairness that exists outside the realm of human flaws of implicit or explicit bias. Many hear of computer programs making decisions in fields with known prejudice such as police enforcement and think that it is an improvement, but many times this is not the case. The belief and trust in these systems with biased results can further increase this propensity. Other times problems of bias can arise in places most would not expect and therefore have no defense against. One example is internet search results in which arrest related advertisements are much more likely to appear for searches of black names. In terms of gender bias, searches of particular jobs such as "CEO", often return very sexist google image results. Another example is simply the naming of AI systems. The most ubiquitous and popular service programs such as Amazon's Alexa or Apple's Siri are personified as female, and their role is do our bidding . A risk that could arise from these examples is "stereotype threat," in which the idea of a stereotype associated with a characteristic of an

individual can lead to a self-fulfilling prophecy (Steele, 2010). The rise of machine learning is riddled with issues of bias and because of its expedited implementation, we do not currently have many systems in place to detect or more importantly, prevent this.

Body

Gendered AI Systems

Siri, Alexa, Watson. Most people have heard these names and know exactly what they do. Siri was the beginning of cellular personal assistants, Alexa has taken over people's homes to do our bidding, and Watson was part of the start of machine learning, a super computer built for the sole purpose of winning Jeopardy. They are some of the most well-known examples of AI and that is partially because of how we have personified and interacted with them. We think of them as more than systems and machines because of their names and how they seem to stretch the limits of what we think computers can do. However, the names chosen for their purpose perpetuate sexist ideals. Siri and Alexa are both systems designed to serve us. We say the name and a command, and the female system comes to life from her patient waiting mode and does her best to grant our wishes, often offering apologies if unable to assist or find something. Watson was made to beat the smartest guys out there and answer any question posed to him whenever he can and wants to (Markoff, 2011). The most iconic and known examples of AI are portrayals of sexist stereotypes and nobody seems to notice or question it. That is the secret with AI. It is close enough to humanity for us to personify and gender it, but far enough that we don't think the system could possibly have humanity's flaws.

Another example of this personification is the chatbot, Julia (Zdenek, 1999). She was created to be indistinguishable from a living person in conversation as an entry of the annual

Loebner Prize, which she got third place in 1993 (Loebner Prize 1994). Every year chatbots from all over the world compete and are judged based on how real and lifelike their program is in the tradition of what Alan Turing called the imitation game. Developers created Julia to be a secretary of a chatbot space, and therefore coded her to interact with people in conversation, enforce rules, and most commonly react to insults or sexual advances. Many of the responses are largely unanimous when similar statements are fed in. One of these common responses is "I'm PMSing," a catchall phrase for when a woman is confused, emotional or wrong, and an easy hack for the most likely male creators of Julia to cover a lot of ground. Though it is an easy solution that adds a biological and therefore lifelike quality to a lifeless system, it perpetuates sexist ideas that women have a natural imbalance that causes them to act more irrationally than men. It is inaccurate and degrading to believe that once a month (or in Julia's case much more commonly) women all lose our ability to use cognition, yet these biases are coded into an award winning chatbot.

Word Associations

In 2013, Google researchers started a word embeddings project to make a neural network system that processed the semantics and usage of words. It was fed in millions of Google News articles as the basis for its data, a set that should be a straightforward, unbiased data set of American society and how we use language. The system would analyze the words for patterns and essentially create a map of these words with weighted connections based on their semantic connection strength. The researchers then could understand the system with simple vector algebra, which consists of statements like "wife is to woman as husband is to man" or in the

algebraic notation “wife : woman :: husband : man.” These relationships are the word embeddings.

Things get controversial however, when phrases such as “man is to doctor as mother is to ...” Were presented to the system and it returned “nurse.” Or it was asked “man is to computer programmer as woman is to ...” and replied with “homemaker.” In this case there was no researcher bias. The researchers attempted to use the least biased data available, neutral google news articles, but sexist correlations still became a part of the program. This research is a perfect example of how societal biases are the underlying issue and machine learning just plainly reports it without the sensitivity that a human might try to provide. Some may regard this as a good thing. Point out the flaws, so we know how to improve. However, systems are not nearly this simple majority of the time. Something straightforward like this would usually be informing some larger black box system, which means that we would only have knowledge of the result and have no understanding of how it came to it. In this case, the algorithm returns a biased result based on the data, and is then employed for some other use where bias is less easy to identify.

An example of a larger process we know about using word embeddings specifically is translation systems. Google translate is the most popular and it has been found to use sexist assumptions in its translation to gendered languages (Olson, 2018). One example is Turkish, a gender-neutral language, to English in which the phrase “o bir muhendis” translates to “he is an engineer” while “o bir hemsire” becomes “she is a nurse.” Word embeddings are the cause of flaws like these, and it is implemented in other services like Google search, Netflix, and Spotify as well. We know that these problems arise from straightforward results of analysis of data as unbiased as possible, so this may mean we need to form algorithms not working

straightforwardly from the data, but with some sort of helping hand to be better than our current society.

People/Facial Recognition

A few years ago, a company called Beauty.AI boasted of hosting the first ever beauty contest judged by machine learning. It was going to be worldwide, easy to enter, and come back with results based on the foundations of beauty such as symmetry, clearness, and a trained data set. Thousands of women from all over the world submitted photos to the contest. Of the 44 winners, the majority were white, only a handful were Asian, and only 1 had dark skin. It became worldwide news that a "racist AI" came up with these results. People recognized for once that machine learning can be racist and biased, but the news outlets portrayed the situation as the machine being the racist one, shifting blame away from the real problem sources. A few pointed some fingers at the programmers, but the real underlying problem is still biased datasets. It is true the programmers should have discovered this failure earlier, but the reasons they got this result reveal more universal issues than one team who lacked foresight.

Beauty.AI got this biased result because of the most common reason: they trained the system on a large dataset but did not think about the demographics within the dataset. The most commonly used database for facial recognition is ImageNet. In this dataset, 45% of images come from the United States while China and India together are just 3%, despite making up 36% of the world's population (Zou & Schiebinger, 2018). The majority of the photos used to train were of caucasian and so the system learned to optimize for them, while other races were less represented and therefore less recognized. This imbalance was not reflected in the actual submissions and, due to the biased optimization, led to disproportionate numbers of white winners. The

programmers may have assumed with such a large data set (millions of photos) the program must be learning from all of them and the more the better, but that is not how machine learning works. The program will analyze every single photo and adjust weights, which determine the importance of many factors, including racial differences (Zou & Schiebinger, 2018). If there are many more Caucasian photos, then Caucasian features will be weighted as more important and therefore more beautiful.

The logic of these datasets is what causes bias problems for many facial recognition systems. Another case of this mishap are facial recognition systems that errored by stating people of Asian descent' eyes were blinking. One case was in a passport photo system (Griffiths, 2016) and another is on cameras with systems that notify if the user appears to be blinking (Rose, 2010). A camera that went under a lot of scrutiny for this error was Nikon because it is a Japanese company, so you would expect it to work for Japanese people. The most likely reason, however, is that the programmers for the software component were either white or were keeping white people in mind, using a dataset and test cases of mostly Caucasian faces.

Another incident under a lot of public scrutiny was when a webcam system couldn't detect a black user but had no problem with a white user (Rose, 2010). There have been many reported cases of people with darker skin in darker environments not registering with recognition systems. One black woman even donned a white mask and found that the system was able to pick up her white "face," but it couldn't register her real face (Buolamwini, 2019). With our advances in technology and machine learning, that shouldn't be happening. Machine learning should be able to detect the slightly more subtle difference in light because the general patterns of shapes and edges, what is most needed for object recognition, are all there. It could be that

programmers just haven't put in the extra effort or insight to make sure their product is as usable for everyone as they should. This example can be taken even more extreme in the case of google photos recognition algorithms labeling two African American users as gorillas (Pulliam-Moore, 2015). Google couldn't figure out a quick fix and ended up just taking out "gorilla" as a possible label. There was no quick fix because it came down to the datasets they train on and the lack of forethought from the original programmers. If more attention was paid to the diversity of the data sets, incidents like this wouldn't happen with such frequency. A case that proves that training AI on a more diverse dataset will result in much less biased results was an MIT gendering system that was less than 1% inaccurate for white men, but about 35% inaccurate for black women (Buolamwini, 2019). It incorrectly labeled Oprah Winfrey, Michelle Obama, and Serena Williams as men. After this glaring flaw was pointed out, it was retrained on a dataset of 1,270 images balanced in gender and ethnicity (Zou & Schiebinger, 2018), and that resulted in the inaccuracy rate for black women and white men to both be around 3%, a drastic improvement for black women and a slight worsening for white men (Olson, 2018).

Search Engines

We use search engines to understand the world and learn, so our perceptions of the world will be changed by the information we find. If that information is biased, then our understanding of the world will become biased as well (Howard & Borenstein, 2018). However, there has been repeated criticism and ridicule of how sexist and racist web searches can often be. There were many news sources that commented on how if someone googles "CEO" the first female result is CEO barbie (Butterly, 2015). That particular flaw has been fixed and many others have as well, but web searches have as long a history of bias as web searches go back. Along the same vein as

Barbie CEO, it was found that if you search for "doctor," the result will be nearly all men, and when you search "nurse," the result is nearly all women (Kay et al., 2015). These results come from stereotypes that we have for professions. It may match somewhat with what we expect to see, but it does not line up with society completely. It is true that 90% of nurses are female (Muench et al., 2015), but it is also true that 34% of doctors are female (AAMC, 2016). These demographics were not being portrayed in the results, but can be coded in as requirements. A study from Datta et al. (2015) also found that Google would display many more high-paying executive jobs to men than women. Some may think this is justified because there are more men in high-paying executive jobs, but it goes back to the question of what do we want to achieve with technology? Do we want to replicate and perpetuate current standards and stereotypes of society or make better ones? Some may think that it is not the job of the programmer to decide right and wrong, and that is a job left up to a higher being (Ekström, 2015), but society is becoming fully integrated with technology, and if it is not abiding by any code of ethics and only reflects back the problems, then it is not an aid to society.

Not only will you find biased representation of sexes in search engines, but also fully racist results. When searching for black names online you will likely get links advertising criminal record checks, no matter how much the individual has worked and succeeded (Sweeney, 2013). Names such as DeShawn, Darnell, and Jermaine were found to bring up ads suggestive of arrest on about 90% of occasions while names like Geoffrey, Jill, and Emma resulted in these ads about 25%of the searches, though sometimes none at all. Likewise, if you search for "3 white teenagers," the results are as expected: smiling, happy, suburban teenagers. When you search "3 black teenagers," the results are mugshots (Kennedy, 2017). If anybody tries to argue that this

phenomenon is still a case of reflecting society and therefore okay, I would unfortunately be forced to concur on the reflection of society's implicit biases. However, I think everyone can agree that it is not a fair assessment of individuals and highlights the injustice in society. Harvard professor Latanya Sweeney (2013) conducted the research investigating discriminatory ads after she, an acclaimed professor and researcher, would get arrest checks as her first google result. Obviously, that is not a fair judgement.

Employment

Though there is not much direct evidence of bias occurring in machine learning hiring algorithms, we do know that more than ever programs are involved in the hiring process. They are trained on past hiring data, and there is an extensive shadow of bias from human hiring managers. In a groundbreaking study, resumes were sent to companies, some titled with white sounding names and some with black sounding names (Bertrand & Mullainathan, 2003). The white names were 50 percent more likely to be called in for an interview and when the resumes were higher quality white names received 30 percent more callbacks while black names had a much smaller increase. It is unlikely that any of these companies had been using machine learning while screening applicants since the study occurred in 2003, but it was found in 2016 that 72% of resumes are filtered out by machine learning algorithms before even having the chance of a human seeing it (ResumerterPro). We can hope that measures have been taken to not allow results like the 2003 study, but the track record of programmers is to not have much foresight when the past data is already representing a bias. They may think they are coding to represent what the employers want by feeding in all the past employment data, but unless specific safeguards are coded in, any and all bias will only be maximized.

One company that did come into issues using AI for recruiting was Amazon (Dastin, 2018). From 2014 to 2015 they had used a code to judge applicants' resumes and suggested recruiters had leaned on it heavily. Then it was discovered that the program had a strong bias against women. The algorithm had learned from a majority men dataset (since mostly men apply) that also mostly men are hired, so it started to penalize applicants with the word "women" on their resume, such as for clubs or women's universities. Amazon scrapped the program once they realized this extreme fault (after it was running for nearly two years) and tried another approach. The program looked for certain keywords that were found on successful resumes in the past such as "executed" or "captured," but it was then argued that those are psychologically masculine words and will therefore lead to more bias. Either way the program didn't work for their intention at all because it sent back resumes that didn't have the proper coding requirements. Nihar Shah, a professor at Carnegie Mellon researching machine learning, commented that "how to ensure that the algorithm is fair, how to make sure the algorithm is really interpretable and explainable – that's still quite far off" (2018). A lot more work needs to be done to create unbiased machine learning in all fields, yet it is still being rushed into sectors like medicine and the justice system as an end-all solution where it can lead to even more harm.

Medicine

The problems that can arise with the intersection of machine learning and medicine is once again caused by biases present in data and non-diversified data sets. Researchers investigate disease and cures through cutting edge technology and nowadays that involves machine learning. Researchers recently were investigating skin cancer from photographs using machine learning to aid in diagnosis (Zou & Schiebinger, 2018). They trained their model on a data set of over

100,000 images with 60% from google images. However, fewer than 5% of these images contained dark-skinned individuals. With the percentage of dark skin so minimal, most machine learning programs treat individuals with dark skin as a minimal population and optimize for the light skinned majority. This bias occurs by diminishing the weights of factors useful for dark skin cancer diagnosis and increasing the weights for light skin cancer diagnosis.

In 2018, doctors and researchers concerned for the future of medicine outlined possible bias that may arise because of machine learning concerning the treatment of patients (Rajkomar et al.) One worry is about current projects that aim to monitor patients for deterioration and signal for immediate intensive care. When building the model, researchers use past medical historical records, but they are concerned the racial demographics of the dataset won't create a universally accurate result. They are aware of the biases in other fields based on the unequal training data, and emphasize that if the same errors occur here, then people's lives are on the line. The system may not alert when it should or erroneously too often, causing a "boy who cried wolf" effect, putting lives at risk. Doctors and researchers attempted to put their needs for these systems out in the world through this article. However, the troubling factor is that it is a restricted medical research article where few, if any, programmers who develop the systems will see it. We can hope that there is input from the users involved going into the development, but the pattern for most of machine learning systems is that they are created by private companies and then sold to the users with little information of how or why they get their results (Howard, 2017).

Law and Justice

Machine learning is being incorporated into our police and justice systems at an increasing rate. There have always been criticisms of bias by individuals within the system so some may think that incorporating a uniform and outsider computer system is the best solution. However, once again because the data is already biased these systems can and have perpetuated biased aspects of the justice system.

Child protective agencies.

The child protective agencies of the US have been overburdened by reports of families to investigate and it has been left to the discretion of each agency to determine who are the most high-risk children to follow up with (Chouldechova, 2018). Machine learning algorithms were a solution turned to since there are high volumes of documentation and data, so predictive risk models were established to determine which cases were most likely riskiest. However, racial disparity is a widely acknowledged problem of the child welfare system currently because of disproportionate need among families of color, geographic context, and implicit and explicit racial discrimination by child welfare professionals. Models trained on this data will only exacerbate the issue through more calls and data collected on minority families.

Predictive policing.

A similar example of this bias is in the case of criminal justice. We know that criminal data is biased because it is not a measure of crime truly, just when it is reported. Sometimes crimes as minor as walking on the street when there is a sidewalk are reported. Contrastingly, it was also found by the Department of Justice that 52 percent of violent crimes went unreported from 2006-2010 (Truman & Langton, 2007). All of this data is then what gets fed into predictive policing algorithms. They are computer programs that forecast where future crimes are most

likely to occur, whether it is specific geographic areas or specific people. There are not exact facts known about these systems because the companies that own them keep their algorithms and inner workings a secret. Despite this secrecy, research has been able to determine that the systems lead to more enforcement in communities that are already heavily, and arguably overly, policed. The systems usually are trained on past police data such as where crimes have been reported, which is known to be biased against black or lower income communities, and the systems tell the police to visit these areas more and report more crime in those areas. These programs aren't being impeded or questioned either. Of the 50 largest police forces, 31 have either used or plan to use one of these systems.

The programs fall under two categories: in one, areas and times are forecasted to have crime occur and are called "hot boxes," and then in the other people are forecasted to either commit crimes or be a victim. The people forecast program is particularly interesting because they usually utilize a social network analysis, where data of a person's criminal past and that of people they know are included. In this network if you are affiliated with many or any people who are likely to commit crimes then your likelihood of threat is increased. In a more extreme case of extensive data use, one system called "Beware" gives each individual a threat score based on data from commercial data brokers. There is no research or findings into how accurate or what exactly these forecast programs use to measure, which is why there could be a major problem of bias that we may never know about.

Furthermore, it was found that the systems predict crime, but give no recommendation or standard on how to use it. So rather than using knowledge of possible crimes to focus on programs that prevent the crimes, police have just focused on more citations and arrests,

particularly for minor crimes that occur in these "hot box" areas. If the programs were used for civic engagement and improvement, rather than more arrests and criminal reporting, there might be more value in them. Then police won't be increasing their odds of visiting a place repeatedly with every arrest they make and subsequently creating more biased data for the program to evaluate. Instead, they could be making a lasting improvement to the people and places that need it.

Risk assessment scores.

After arrest, defendants could still have machine learning unfairly judge them. Machine learning programs have been created to predict the likelihood of a person re-offending by assigning a person a risk assessment score (Angwin et al., 2019). Propublica, an investigative journalism team focused on exposing abuses of power, investigated the COMPAS scoring systems after U.S. Attorney General Eric Holder raised concerns in 2014, suggesting that risk assessment scoring systems like COMPAS could "inadvertently undermine our efforts to ensure individualized and equal justice" and "may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society." This sentiment may sound familiar after all the troubling examples I have previously stated. The Propublica's results supported Holder's concerns by demonstrating significant racial disparities. The mistakes were near the same rate for black and white defendants, but in opposite directions. The program inaccurately labeled black people as likely to recommit crime at twice the rate as white defendants, and white people were inaccurately labeled as not unlikely to recommit crime much more than black defendants. The researchers investigated if the disparity could be because of the person's record or type of crime committed by running a statistical test isolating the effect of

race from criminal history, recidivism, age, and gender. In this separated analysis, black defendants were labeled 77 percent more likely to commit a future violent crime and 45 percent more likely to commit a future crime of any kind. These results are partially because of extreme racial bias within the criminal record of our country without machine learning, but sometimes the systems draw from indirect sources.

One of the most popular products of this kind is from Northpointe, which checks criminal records in addition to asking 137 questions such as "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" or if "a hungry person has a right to steal." The company claims that they never ask about race and therefore are not racially biased, but in our society African-Americans and Hispanics make up 32% of the US population, but 56% of incarcerated people (NAACP, 2019), so therefore it would be more likely for these minority groups to have had a parent in jail. In addition, minority racial groups are more likely to experience multidimensional poverty including low household income, limited education, no health insurance, low-income area, and unemployment than white counterparts (Reeves, Rodrigue, & Kneebone, 2016) which leads to food insecurity (USDA & U.S. Census Bureau, 2018) and is a factor associated with drug abuse (Chaloupka et al., 1999). There are associations to race deep-rooted in our society and the data reflects that, which is all these machine learning systems have to go off of.

Facial recognition for criminality.

One of the more ridiculous sounding applications of machine learning to criminal justice is a neural network that claims to identify criminals by their faces (ArXiv, 2016). There is a history of this idea in the theory of Cesare Lombroso, an early criminologist, who suggested that

criminality is a quality we are biologically born with and can be determined by features such as a sloping forehead, large ears, and various asymmetries. He was statistically proven wrong by criminologist, Charles Goring, who compared physical features of criminals and non-criminals. However, in 2011, a group of psychologists from Cornell found that people were quite good at distinguishing criminals from only photos, bringing the debate back to light. The neural network brought up here is the work of two Chinese researchers, Xiaolin Wu and Xi Zhang who trained their data on Chinese men and then correctly labeled criminals at an accuracy of 89.5%. They even identified certain features of the face as ones that contributed to the result such as the curvature of the upper lip and distance between the inner corners of eyes. They simplified a description of the faces of criminals as usually much more dissimilar or out of the ordinary than the more common resemblance of non-criminals.

That fact raises some interesting questions such as if criminology is caused by a shunning or unfairness of society based on physical appearance? Do people fall into crime because of a stereotyped appearance? Are people of minority ethnographic backgrounds with varying facial structures targeted in arrests in China as African-Americans are in the U.S.? The final question enters into the dangers of research like this. If researchers had developed a similar idea and project in the U.S. and claimed to predict criminality from the U.S.'s criminal records it would likely be outlandishly racist because of the racism existing in the data. In China's case, it is mostly homogenous with 90 percent of citizens belonging to the Han Chinese group. The others, who are different ethnic backgrounds, foreigners, or mixed race, have been discriminated against (Internations). This discrimination could likely extend to the rate at which crimes are reported

and pursued similar to the U.S. and that could contribute to the varying face hypothesis found in Wu's and Zhang's research.

Discussion

Possible Future Bias

Robot police.

There are plans and a few implementations of police robots being utilized in the future (Howard, 2017). Again, some may hear that and think it's a good idea, better than the current racism and bias present in crime enforcement. However, the bias is being coded in and can be just as bad or worse in these robots than current police. In 2015, North Dakota became the first state to allow armed drones (Wagner, 2015). Though the original bill was supposed to allow unarmed drones, the final decision allowed non-lethal weapons like tasers and rubber bullets. It is unclear if the drones are automated currently, but it is only a matter of time with the current progression of machine learning being integrated into the criminal justice system. There has already been a recorded case of police robot having a direct role in the death of a black man in Dallas (McFarland, 2016). Though not many details were given, the robot was sent to detonate a bomb near a subject who had killed 5 police officers. It is most likely the robot was not AI and had a human controller, but it is an example of a shift in policing. It was the first case of an American citizen being killed by a police robot. If it is done once why not in the future? And why not when machine learning is controlling it rather than a human?

Self-driving cars.

Another technology that will likely be everywhere soon is self-driving cars (Howard, 2017). There are a few models with features of it available, but as we progress further with these

autonomous cars, computers will be literally taking the wheel making life or death decisions. The programs must address many moral decisions that we may not consider a part of driving, but will and have occurred such as should the car swerve to avoid hitting an animal if that puts the driver's life at risk? What about a child? What about an elderly person? Questions such as these are in shaky moral territory where people don't want programmers being the only ones making decisions. To remedy this, MIT put together a site, moralmachine.edu, where people can go on and judge certain situations about what they think the best decision is (Moral Machine). Some scenarios go into extreme depth such as should a self-driving car in which the brakes fail swerve into a concrete barrier, killing 2 homeless people and 2 men, or not swerve and plow into a pedestrian crosswalk where 2 male executives, 1 female executive, and 1 woman are illegally crossing and would be killed. Being presented with a complex situation like this makes me wonder how this data will be used. Why does asking about "executives" matter because how would the machine identify pedestrians as "executives?" Will it make a machine learning decision based on the appearance of the pedestrian? If that is the case, then programs that go into judging people based on appearance will go into effect with lethal consequences. I have repeatedly explained research and opinions about how machines learning is nowhere near being an unbiased resource, so any programs judging people (particularly in the case for if they are "executive," in which it has already been shown that machine learning algorithms target white men as this) will be biased with utmost certainty.

Even if there are somehow not directly biased machine learning decisions coded in, it has also been exposed that object recognition software in autonomous vehicles do not register people with darker skin as well as people with lighter skin (Wilson et al., 2019). Some people had

suggested the reason for this is that data from night hours makes it more difficult for systems to see darker skin as any human would too (which is still troubling), but the researchers found no significant difference between day and night conditions, refuting that belief. What the researchers did find, however, was that the dataset these object recognition systems are trained on include 3.5 times more light skinned people than dark skinned people and all of the data is weighted equally. The researchers suggested that if data scientists broke up the training with separate trainings and then a weighted equation between those, there could be better optimization for all skin types, rather than only optimization for the light-skinned majority.

Solutions

I have outlined many fields and reasons for how and why machine learning can be biased against groups of people but will now delve into ways this bias can be prevented. It depends on the reasons for the bias, so I will focus on the most common reasons and most applicable solutions.

Diverse business teams.

In terms of the anthropomorphizing of systems into certain genders, we need intelligent business and programming decisions that don't perpetuate sexism. Instead of naming your digital servant a female name, make it gender neutral, a made-up word, or let people name their own. For sectors of life where there has been perpetual sexism like in assistant/service jobs there needs to be more thought about the landscape of the field. In this particular instance it most likely goes to the business executives who decide how to market their product. If there is more representation of women and minorities in these companies then it is much more likely faults like this wouldn't be overlooked. Technology is a field where much more knowledge and

concern about the sociological effects of decisions should be cultivated since technology is being weaved into every other field of society.

Data related solutions.

For the bias that occurs because of data, there are a few approaches that can be taken. As I've stated before one of the most frequent and easiest solutions is to make the data set more balanced, but to do that we need to have information about the demographics of data sets. Some suggest that the users of programs should be more involved in the design process. Others just want more transparency of how algorithms come to the decisions they do. At the very least there is a call for more case-based testing of programs focusing on possible biased outcomes becoming a standard operating procedure.

Biased data exclusion error.

One more obvious suggestion is to just leave out the information in datasets that says ethnicity, gender, or anything else that could lead to a biased outcome. The problem with this approach is that it has been proven that even if direct indicators of bias are removed the other data will still express the bias (Redreshi et al., 2008). For example if race is removed, the data could still have where the person lives, their income, their criminal record, their employment record, and other socioeconomic factors, which all are influenced by race and effectively will result in the same bias even without having the direct reason why in the data. Another somewhat debunked approach is to have programmers test their data by seeing the result of their code with sensitive data like race or gender and then compare it to the result if that sensitive information is removed (Hardt et al., 2006). If it's the same percentage supposedly it is not biased. Though it is

less likely the program is biased, there still could be bias either way with reason similar to the idea of leaving information out, the other information available leads to the same bias even without the definitive datapoint.

Balanced data sets.

There are some researchers who think the solution to having unbalanced datasets is to make it standard for datasets to have a "nutrition label" that has an overview of the "ingredients" present in the dataset (Holland et al., 2018). They extol the positives this sort of system could bring such as allowing more robust practices, better allow programmers to choose their dataset, and of course the improvement of the AI programs themselves because then they can be trained on better suited and diverse datasets. The only real negative is that datasets include very diverse items so it may be difficult to come up with a standardized practice that applies to everything. In my opinion, any sort of label giving insight into the variances that make up a dataset would be more than welcome compared to many current datasets, with labels only listing the overall subject. With labels like this, data scientists could better avoid training their programs on biased data and having a label with a breakdown of demographics could make them think about bias when they wouldn't have without the "nutrition facts."

Fair classification formula.

Researchers from Google found another solution (Jiang & Nachum, 2019). They developed a procedure for re-weighting the data that can take into account biases and effectively create a new unbiased dataset for programmers to run their machine learning algorithms on. They start with the assumption that there is an unknown and unbiased label function and a dataset that includes biased data, but was created with the intention of accuracy and without bias.

They are able to represent the bias in the data as a closed form expression and use it to readjust the weights of data to represent a "true" equality. They tested their approach on various standard machine learning fairness datasets and theirs outperformed other methods for fair classification. If this method is learned and used amongst the future programmers of AI, then there might possibly be a simple systematic way of preventing bias in future machine learning tasks, but more research into this method's effects and effective usages is needed to know for sure.

User Focused Programming.

Researchers have also developed a new method for machine learning algorithms in which the emphasis of undesirable results can be factored in by the user (Thomas et al., 2017). Instead of the program automatically minimizing overall error, the user in this case can enter in their own "probabilistic constraints" that must be met. The especially ingenious aspect of this method is that the people who understand the field that the program is being applied to are able to have some power in what sort of results they are looking for. This is in contrast to usual algorithms where the relatively uneducated (in the particular field of utility) programmers are determining what results an application should produce. With applications built in this format, bias could be decreased by the expert field users by being able to go in and make nuanced decisions, ones that might not occur to data scientists, so that the code does not result in troubling results. If this full fledge change is not possible, then a smaller step towards it could be merely having the users of programs more involved in the design process, making sure their values and expectations are being coded. That is the idea behind the MIT self-driving car morality website mentioned previously, which is definitely an improvement from only programmers making those decisions, despite other flaws.

One study that analyzed what factors should be focused on in the design of socially assistive robots for adults diagnosed with depression and other co-occurring illnesses (Šabanović, 2015) is an example of having users' input play a role in the design. A reason these researchers stated for deciding to investigate in this fashion is that of patient-centered care. Because this is a medically related application, the concern for humans was more of a focus for the designers, but maybe more fields should be using similar directions of thought since nearly all technology has a major impact on human lives.

Transparency of Algorithm.

A repeated problem for many of these machine learning programs is that the user would only know the result of an algorithm and have no idea how it came to that result. One particularly harmful example of this is the case of AI's that give risk scores to judges that was criticized by ProPublica (Angwin et al., 2019). The judges receive a risk score about an individual based on nearly every statistic there is on the person, so the judge won't know what exactly causes the risk score. For example, in most of these algorithms if the person is younger they receive a higher score because they just became an adult and are already offending, but what about the case of an 18 year old who is charged with sexual assault of his 16 year old girlfriend due to age of consent laws? I would not think this boy is a hardened criminal, but sexual assaults are taken seriously by these machines (Dressel & Farid, 2018), so would the data points of young and sexual offender make his risk score through the roof? Who would ever know?

There are a few aspects of opacity that arise from the business of AI (Bousquet, 2018). First, these companies usually want to keep their algorithms a secret. The algorithm is their

product so there is some merit in that, but some key elements could be disclosed: what data is used and why, what analytic tools are used on the data, and performance data of the algorithm before and after implementation. To address the issue of people not understanding the technicalities of machine learning code, others want visual representations or simply that only understandable models are used for public matters. A postdoctoral fellow at the Center for Computation and Society at Harvard is quoted supporting this, stating "you can create something you can explain to a policymaker that works just as well as this other stuff." In conclusion, there are many ways that machine learning development can be less of a secluded space and become one that welcomes and is fair to all in development and usage, but if no changes occur to the standards of machine learning development then there could be serious societal detriments.

Works Cited

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2019, March 09). Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ArXiv, E. T. (2016, November 24). A deep-learning machine was trained to spot criminals by looking at mugshots. Retrieved from https://www.technologyreview.com/s/602955/neural-network-learns-to-identify-criminals-by-their-faces/

Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. doi:10.3386/w9873

Bousquet, C. (2018, August 31). Algorithmic Fairness: Tackling Bias in City Algorithms. Retrieved from https://datasmart.ash.harvard.edu/news/article/algorithmic-fairness-tackling-bias-city-algorithms

Buolamwini, J. (2019, February 07). Artificial Intelligence Has a Racial and Gender Bias Problem. Retrieved from http://time.com/5520558/artificial-intelligence-racial-gender-bias/

Butterly, A. (2015, April 16). Google Image search for CEO has Barbie as first female result - BBC Newsbeat. Retrieved from http://www.bbc.co.uk/newsbeat/article/32332603/google-image-search-for-ceo-has-barbie-as-first-female-result

Chaloupka, F., Grossman, M., Bickel, W., & Saffer, H. (1999). The Economic Analysis of Substance Use and Abuse: An Integration of Econometrics and Behavioral Economic Research. *University of Chicago Press,*327-368. doi:0-262-10047-2

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research,81*. doi:81:134-148, 2018

Criminal Justice Fact Sheet. (2019). Retrieved from https://www.naacp.org/criminal-justice-fact-sheet/

Dallas, B. T. (2016, July 11). Robot's role in killing Dallas shooter is a first. Retrieved from https://money.cnn.com/2016/07/08/technology/dallas-robot-death/index.html

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

Discrimination and Racism in China. (n.d.). Retrieved from https://www.internations.org/china-expats/guide/29460-safety-security/discrimination-and-racism-in-china-17752

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances,4*(1). doi:10.1126/sciadv.aao5580

Ekström, A. (2015, November 10). The Moral Bias Behind Your search Results. Retrieved from https://www.ted.com/talks/andreas_ekstrom_the_moral_bias_behind_your_search_results/up-next?language=en

Food Insecurity and Poverty in the United States: Findings from the USDA and U.S. Census Bureau. (2018). Retrieved from https://hungerandhealth.feedingamerica.org/wp-content/uploads/2018/10/Food-Insecurity-Poverty-Brief_2018.pdf

Griffiths, J. (2016, December 09). New Zealand passport robot thinks this Asian man's eyes are closed. Retrieved from https://www.cnn.com/2016/12/07/asia/new-zealand-passport-robot-asian-trnd/index.html

Guynn, J. (2015, July 01). Google Photos labeled black people 'gorillas'. Retrieved from https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/

Hall, W. J., Chapman, M. V., Lee, K. M., Merino, Y. M., Thomas, T. W., Payne, B. K., . . . Coyne-Beasley, T. (2015). Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *American Journal of Public Health,105*(12), 2588-2588. doi:10.2105/ajph.2015.302903a

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *ArXiv*. Retrieved from https://arxiv.org/pdf/1610.02413.pdf.

Holland, S., Hosney, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *ArXiv*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1805/1805.03677.pdf.

Howard, A., & Borenstein, J. (2017). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics,24*(5), 1521-1536. doi:10.1007/s11948-017-9975-2

Jiang, H., & Nachum, O. (2019). Identifying and Correcting Label Bias in Machine Learning. *ArXiv*. Retrieved from https://arxiv.org/pdf/1901.04966v1.pdf.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*. doi:10.1145/2702123.2702520

Kennedy, R. (2017). Algorithms and the Rule of Law. *Legal Information Management,17*(03), 170-172. doi:10.1017/s1472669617000342

Loebner Prize Competition 93: Julia's log. (1994). Retrieved from http://www.lazytd.com/lti/julia/contest93.html

Markoff, J. (2011, February 16). Computer Wins on 'Jeopardy!': Trivial, It's Not. Retrieved from https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

Meserole, C. (2018, October 18). What is machine learning? Retrieved from https://www.brookings.edu/research/what-is-machine-learning/

Moral Machine. (n.d.). Retrieved from http://moralmachine.mit.edu/

Muench, U., Sindelar, J., Busch, S. H., & Buerhaus, P. I. (2015). Salary Differences Between Male and Female Registered Nurses in the United States. *Jama,313*(12), 1265. doi:10.1001/jama.2015.1487

Olson, P. (2018, February 27). Racist, Sexist AI Could Be A Bigger Problem Than Lost Jobs. Retrieved from https://www.forbes.com/sites/parmyolson/2018/02/26/artificial-intelligence-ai-bias-google/#7af6d1ec1a01

Olson, P. (2018, February 19). The Algorithm That Helped Google Translate Become Sexist. Retrieved from https://www.forbes.com/sites/parmyolson/2018/02/15/the-algorithm-that-helped-google-translate-become-sexist/#5b2821e37daa

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*. doi:10.1145/1401890.1401959

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine,169*(12), 866. doi:10.7326/m18-1990

ResumerterPro. (2016, February 16). 72% of Resumes are Never Seen by Employers. Retrieved from https://www.accesswire.com/436847/72-of-Resumes-are-Never-Seen-by-Employers

Robinson, D., & Koepke, L. (2016). Stuck in a Pattern. *Upturn,1*(2). Retrieved from https://centerformediajustice.org/wp-content/uploads/2016/08/Upturn_-_Stuck_In_a_Pattern_v.1.01.pdf.

Rose, A. (2010, January 22). Are Face-Detection Cameras Racist? Retrieved from http://content.time.com/time/business/article/0,8599,1954643,00.html

Steele, A. (2017, September 30). Stereotype Threat. Retrieved from https://www.learning-theories.com/stereotype-threat-steele-aronson.html

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *SSRN Electronic Journal*. doi:10.2139/ssrn.2208240

Thomas, P., Castro da Silva, B., Barto, A., & Brunskill, E. (2017). On Ensuring that Intelligent Machines Are Well-Behaved. *ArXiv*. Retrieved from https://arxiv.org/abs/1708.05448.

Truman, J., & Langton, L. (2007). National Crime Victimization Survey, 2005. *ICPSR Data Holdings*. doi:10.3886/icpsr04451.v1

Wagner, L. (2015, August 27). North Dakota Legalizes Armed Police Drones. Retrieved from https://www.npr.org/sections/thetwo-way/2015/08/27/435301160/north-dakota-legalizes-armed-police-drones

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive Inequity in Object Detection. *ArXiv*. Retrieved from https://arxiv.org/pdf/1902.11097.pdf.

Zdenek, S. (1999). Rising Up from the MUD: Inscribing Gender in Software Design. *Discourse & Society,10*(3), 379-409. doi:10.1177/0957926599010003005

Zou, J., & Schiebinger, L. (2018, July 18). AI can be sexist and racist - it's time to make it fair. Retrieved from https://www.nature.com/articles/d41586-018-05707-8

Šabanović, S., Chang, W., Bennett, C. C., Piatt, J. A., & Hakken, D. (2015). A Robot of My Own: Participatory Design of Socially Assistive Robots for Independently Living Older Adults Diagnosed with Depression. *Human Aspects of IT for the Aged Population. Design for Aging Lecture Notes in Computer Science,*104-114. doi:10.1007/978-3-319-20892-3_11