An Overview of Algorithmic Bias in Artificial Intelligence

Nitesh Kartha
William D. Young
Department of Computer Science
University of Texas at Austin

Abstract: Ârtificial Intelligence has grown throughout recent years to become a major part of popular culture and products used by people around the world. However, these systems are not perfect and can in fact contain multiple different biases in their underlying algorithms. In this paper, we provide an overview of the sources of algorithmic bias, a discussion of real-world case studies and their impacts, and a general summary of past attempts to address biases in artificial intelligence such as the General Data Protection Regulation (GDPR), corporate and governmental ethical guidelines, and New York City's Automated Decision System (ADS) Task Force. Specifically, we discuss the COMPAS algorithm used for pretrial assessments, the Facebook ad-delivery algorithm used on its online advertising platform, and a healthcare algorithm used for high-risk care management in the United States.

We conclude that algorithmic bias will only be exacerbated as more systems become automated through artificial intelligence. However, recognizing and calling for the alleviation of biases in current systems as well as approaching the design of automated systems holistically have led to reduced biases. More empirical research is required to fully understand what ways algorithmic bias can consistently be reduced.

Keywords: artificial intelligence, bias, algorithms, deep learning, ethics of artificial intelligence

Introduction

Over the last decade, artificial intelligence has grown to become an over $300 billion industry [1], with A.I. being used from personal assistants such as Siri or Alexa [2], [3] to automated cars such as the Autopilot A.I. in Tesla electric cars [4]. Similarly, the technologies behind artificial intelligence have become more common knowledge

as shown by the growing popularity of terms such as "deep learning," "neural net," "reinforcement learning," etc. (as measured through interest on Google Trends).[5]–[7]

But what exactly is artificial intelligence? A concise definition of artificial intelligence has eluded scientists [8],but most commercial artificial intelligence revolves around machine learning: training an algorithm to make decisions and predictions based of new data. [9], [10]

Recently, however, we as consumers sometimes overly trust artificial intelligence in their decision-making skills and believe them to be more knowledgeable and neutral about a subject than a human expert. [11] In reality, artificial intelligence algorithms are susceptible to algorithmic biases, akin to human biases, that can lead to inaccurate and in some cases illegal outcomes.

But how and what kind of biases can be found in artificial intelligence systems? Below, we list a group of biases that have been explained by the media but is by no means a comprehensive list of all biases that can be present in such systems.

- *Data-driven biases* can occur when the data used to train the AI is skewed in some way. For example, the facial recognition program in Nikon's cameras and HP computers had trouble recognizing Asian and African-American faces respectively, possibly as a result of the dataset used to train the programs not accounting for those ethnicities. [12], [13]

- *Human interaction bias* can occur when user interaction and user biases influence the model itself. This bias is especially present when user behavior is used to train/adapt the model and can be particularly disastrous if proper filtering of behavior and safeguards are not in place. A notable example of this bias gone wrong is Microsoft's Taybot which was a Twitter bot that became racist with neo-Nazi tendencies as a result of user interaction. [12], [14]–[16]

- *Emergent bias* also known as *similarity bias* typically occurs in recommendation systems that focus on "user relevancy." As a result of trying to cater relevant content to users, these algorithms can create "filter bubbles" where opposing or different views are considered irrelevant and therefore never shown to users. The implications of these "filter bubbles" on society are great and are still be researched to this day. [12], [17]

Similarly, bias can occur in three key stages of development [18]:

- *Framing the problem*: Translating societal problems into mathematical problems for algorithms can be challenging and when a business frames a problem for a business goal (i.e. optimizing profits) rather than a holistic goal (i.e. increasing health care coverage), bias can occur from the algorithm even if unintentional.

- *Collecting the data*: Data used for training can be unrepresentative of reality or can reflect existing prejudices which are then reinforced by the algorithm. (i.e. Nikon and HP examples from above)

- *Preparing the data*: Selecting which attributes the algorithm considers (also known as a policy choice) can greatly impact how accurate the algorithm is and how biased it can be.

In this paper, we will examine three cases of prominent algorithmic bias in artificial intelligence–the COMPAS algorithm for determining recidivism, Facebook's ad delivery algorithm, and healthcare algorithms for high-risk health management programs–and their impact on society. We conclude by discussing ways to mitigate algorithmic bias in general.

COMPAS Risk Assessment Algorithm

Since the mid-20th century, techniques have been used to predict crime and recidivism in the United States in order to create an efficient and expedited justice system. [19] As a result of mass incarceration in recent decades, the use of automated solutions for pre-trial and sentencing protocols have widely increased, especially given the opportunity for automated solutions to be less biased than humans judges in the American judicial system. One popular type of automated solution is risk assessment algorithms.

Risk assessment algorithms, more specifically, are defined as algorithms that "purport to predict future behavior by defendants and incarcerated persons" [20] and is typically used in bail and other pre-trial decisions but have also been used in some states for sentencing [21]. Typically, these algorithms use surveys and public records to determine a risk score that can guide a judge to determine how strict a sentence/bail should be for a given defendant. [20]

The appeal of risk assessment algorithms and other forms of predictive policing is that they can provide ways of predicting (and therefore preventing) future crimes. Similarly, they can discourage recidivists from committing more crimes while preventing low-risk criminals from having to go to jail in the first place, reducing prison populations while keeping communities safe. Algorithms can also be more fair at determining sentencing than a human judge and therefore serve as a "neutral arbiter" that makes decisions based on data and not on biases.

However, risk assessment algorithms have been criticized for being opaque in how they make their decisions (akin to a "black box") and therefore it is hard to understand their processes and appeal the judgements they make. In fact, there is currently no procedure in place to appeal assessments/scores made by algorithms in the United States. Algorithms can also have biases like humans, especially when using historical data which can propagate historical biases even further. This leads us to our discussion on the bias of one of the most popular risk assessment algorithms, the Correctional Offender Management Profiling for Alternative Sanctions algorithm or COMPAS for short.

ProPublica Investigation into COMPAS

In May 2016, *ProPublica* released an article underlying their analysis of the COMPAS algorithm in Broward County, Florida [22]. Broward County was chosen because

of its large population (almost 2 million according the the 2019 Census Bureau estimate) as well as Florida's open record laws. [22], [23] *ProPublica* found that the COMPAS algorithm displayed significant racial bias when handing out risk scores, with African Americans being classified as higher risk defendants more often than White Americans. [21]

COMPAS is one of the most widely used risk algorithms in the United States and is used statewide in 4 states and several other jurisdictions. [21] The algorithm was created by Northpointe to be used in pre-trial hearings and determines scores from a 137 question survey and public records. [22] The scores range from 1 to 10 and Northpointe considers scores ranging from 1 to 4 as "low risk" and scores 8 to 10 as "high risk." [22] It is crucial to note that Race was not explicitly asked in the questionnaire or provided to the algorithm which means that any bias that the algorithm generated was implicit [22].

The *ProPublica* report noted that although Race was not explicitly asked, questions such as "Was one of your parents ever sent to jail or prison?" as well as "How many of your friends/acquaintances are taking drugs illegally?" were asked and due to systemic racial inequality, these questions could have led to the bias in the algorithm itself [22].

But what biases did the COMPAS algorithm have and how apparent were they? *ProPublica* analyzed the algorithm using a logistical model that tracked demographic information and risk scores and found that actually the most predictive factor for "high risk" scores was age, not race. Defendants under the age of 25 were "2.5 times more likely" to score higher than middle aged offenders, even after controlling for prior crimes, future criminality, race, and gender. [22]

This bias perhaps makes sense. A 2017 study by the *United States Sentencing Commission*, the federal independent agency tasked with reducing sentencing disparities and promoting transparency in sentencing, found that younger federal offenders (younger than the age of 21) were over 50% more likely to recidivate.[24] As a result, COMPAS's bias on age can be used as an example where bias could be inherent as part of the goals of algorithm itself (i.e. if you want to reduce recidivism and younger people are more likely to recidivate, you should rate younger people with a higher risk of recidivsm...whether or not they actually do) and is not necessarily discriminatory. However it is important to note that since the bias exists after controlling for prior crimes, future criminality, race, and gender, this means that younger defendants who will not recidivate are likely to be rated higher risk anyway.

More controversially and the bias focused on by *ProPublica* was the use of race as a predictive factor for recidivism, despite the algorithm not having access to racial

data directly. Black defendants were "45% more likely" to get a higher risk score than white defendants and "77.3%" more likely to receive a higher score for violent recidivism (crimes such as "murder, manslaughter, forcible rape, robbery," etc.) than white defendants even after controlling for criminal history and future violent recidivism. [22]

Accuracy of COMPAS

As a result of *ProPublica* report, it begs the question whether COMPAS is actually accurate at predicting/measuring risk of recidivism? After all, if COMPAS is equally accurate among all races, it could simply suggest a correlation between race and crime present in the sample used to calibrate the algorithm or in the population of Broward County (although it is imperative to note that multiple studies have found no such correlation between race and crime despite the assumption made in criminological literature [25]).

ProPublica found that accuracy was consistent between races (62.5% of white defendants who was predicted to recidivate did so within 2 years of being scored while 62.3% of black defendants predicted did so within 2 years as well) but lower than the threshold that NorthPointe described for "satisfactory predictive accuracy" which was 70% or greater. [22] The predictions made by COMPAS for violent recidivism were only 20% accurate. [22]

Another finding by *ProPublica* was that high-risk white defendants were 3.51 times more likely to recidivate than low-risk white defendants while high-risk black defendants were only 2.99 times more likely to recidivate as low-risk black defendants [22]. This suggests that the spectrum of scores are not as valuable for black defendants as there is less disparity between those scored as "high risk" vs those scored as "low risk" (albeit by a slightly lower margin than white defendants).

Perhaps the most controversial racial disparity was that found in false positive rates. COMPAS was found to misclassify a black defendant as higher risk almost twice as often as white defendants (45% false positive rate vs 23% false positive rate) [22] which given its use in pretrial and sentencing decisions could have led to harsher sentences for African Americans than justified by the risk of recidivism. COMPAS was also found to underestimate white recidivism greater than black recidivsm (48% false negative vs 28% false negative) [22], suggesting that racial bias did in fact affect the overall accuracy of the algorithm.

Implications of COMPAS and Risk Assessment Algorithms

Although *ProPublica*'s report and analysis of COMPAS was not the first conducted, it appears to be the only one to analyze any possible racial bias in the algorithm. New York State began to use COMPAS in a pilot program in 2001 and then rolled out the algorithm throughout the state's probation departments (besides NYC) in 2010 but only released a statistical evaluation in 2012 which only included findings about overall accuracy and no analysis regarding bias [21]. In response to questioning about why the state did not analyze racial bias, the spokeswoman for the State said that the study "only sought to test whether the tool had been properly calibrated to fit New York's probation population" and noted that "nearly all New York counties are given [COMPAS] assessments during sentencing" [21].

Similarly, Northpointe themselves conducted a study on COMPAS in 2009 and found that the algorithm was "slightly less predictive for black men than white men" but didn't examine any other racial disparities beyond that. [21] In fact, even Wisconsin, a state that in 2012 began using COMPAS at "every step in the prison system, from sentencing to parole" had not conducted a study of the tool and continues to use the scores at "every decision point." [21]

The use of COMPAS in Wisconsin had become so controversial that it led to the landmark court case *Loomis v. Wisconsin* which can be used as a model of what limits the judiciary currently believe are necessary for risk assessment algorithms as well as where they should be used.

Specifically, *Loomis v. Wisconsin* was a court cased ruled by the Wisconsin Supreme Court in 2016 regarding a black defendant Eric Loomis who argued that the use of COMPAS in sentencing decisions deprived people of their due process rights since it considered gender in the assessment and due to the secrecy around the methodology of how COMPAS determines a risk score. [26]

The Wisconsin Supreme Court however ruled in favor of the State of Wisconsin and the use of COMPAS, finding the argument against the use of gender unfounded as gender was used for accuracy purposes and not necessarily considered as a criteria for sentencing. [26] The court also emphasized the idea that COMPAS should be used as part of many factors when it comes to sentencing and prescribed the ways in which such assessments should be used [26]:

- Assessments cannot be used to determine whether an offender is to be incarcerated

- Assessments cannot be used to determine the severity of the sentence (a very confusing point considering that COMPAS is used by many judges to determine what kind of sentence a defendant should receive [21])
- COMPAS specifically must include written warnings for judges that emphasize the following:
  - The proprietary nature of the algorithm
  - COMPAS's inability to identify high-risk individuals due to group data
  - The fact that there is no cross-validation study between the national sample used to calibrate COMPAS and the population of Wisconsin.
  - COMPAS might disproportionately classify minority offenders as having higher risk of recidivism
  - COMPAS was created for post-sentencing determinations.

Although these warnings seem logical to include, the fact that they are mandated by the Wisconsin Supreme Court begs the question as to why risk assessments like COMPAS are still legally allowed to be used for sentencing? The *Harvard Law Review*, in their analysis of the *Loomis v. Wisconsin* case, notes that "written disclaimers [required for assessments like COMPAS] is unlikely to [enable judges] to better assess the accuracy" of the algorithms and the "appropriate weight" that such assessments should have while making judicial decisions [26].

The *Review* also noted that "encouraging judicial skepticism" of algorithms like COMPAS alone "does little to tell judges how much to discount these assessments" [26], and that "[s]cholars warn that these assessments often disguise...overt discrimination based on demographics and socioeconomic status" [26].

This problem is not necessarily unique to COMPAS or even automated risk assessments in general. A 2010 study found that "risk" as defined through risk assessments usually become a proxy for race since the calculation for recidivism risk is heavily dependent on prior criminal history and because of systemic racism in the criminal justice system, prior criminal history usually becomes a proxy for race [19].

Similarly, since the 1970s, when it no longer was politically viable to use race directly in risk assessments, states adopted actuarial models that use less predictive factors (ranging from around 25 factors in the 1930s to roughly 5 by the 1990s) and focused more on prior arrests, convictions, and incarcerations despite the fact that it is unclear whether prior criminal activity actually works to predict risk of recidivism. [19] Similarly because of continuously increasing racial disproportionality in the prison population, the focus on prior criminality hits African American

communities the hardest which suggests that risk assessments like COMPAS have a long way to go before becoming the "neutral arbiter" that proponents of predictive policing technology desire it to be [19].

Facebook's Ad Delivery Algorithm

In 2020, online advertising in the United States has grown to be a $120.9 Billion industry, greater than all other forms of advertising combined and projected to grow even more in coming years. [27] However, unlike traditional advertising where content creators and companies can decide exactly how and when a consumer sees their ad, online advertising instead leverages algorithms to conduct immediate auctions to determine which ads are shown and to whom. [28]

Facebook specifically is one of the largest advertising platforms in the world, home to at least 8 million advertisers and 2.6 billion monthly users on Facebook (and 1 billion monthly users on Instagram). [29] Advertising generated $69.7 billion for the company in 2019, more that 98% of its total revenue. [29] As a result, Facebook has developed many tools for content developers to deliver relevant ads/ads that are more likely to be interacted with, mostly through the use of algorithms and artificial intelligence. [28]

However, like with COMPAS and risk assessment algorithms, online advertising algorithms are very opaque and aren't available to the public due to trade secret laws. As a result, it is hard to prosecute advertising that is discriminatory (which can be illegal in certain cases in the US), and harder still to understand how bias could arise even when content creators do not explicitly target certain audiences.

Northeastern University Study

In March 2019, the US Department of Housing and Urban Development sued Facebook for letting advertisers target their ads by race, gender, and religion which are considered *protected classes* under US law. Although, Facebook claimed to restrict the targeting parameters so that advertisers cannot explicitly target ads based off those classes, researchers at Northeastern University found that the ad-delivery algorithm used by Facebook carries out the same discrimination using demographic information. [30]

Before we delve into what the researchers at Northeastern University discovered about Facebook's ad delivery algorithm, it is important to clarify what exactly that

means. There are two phases to online advertising platforms: *ad creation*, where advertisers (also known as content creators) submit text and images that make up the content of their ad and targeting parameters for who sees the ad, and *ad delivery*, where the platform delivers the ads to specific users based on a number of factors. [28]

The researchers noted that although Facebook did limit the explicit targeting features to prevent discrimination, the ad delivery process itself led to skewed results. Typically, platforms desire to show user's "relevant ads" (ads that particular users are more likely to engage with) even though advertisers typically don't know "a priori" which users are more receptive to their advertising. [28], [31] This is usually done through creating extensive user profiles and tracking ad performance but this historical data can allow for ads to be delivered to a skewed subgroup of the population. [28] This is especially worrying when done to advertising about credit, housing, and employment which is protected under federal law in the United States. [28]

Similarly, the researchers noted that market effects and profit-maximization strategies can play a role in causing skewed ad delivery. For instance, it is commonly known that certain users (typically users around 18-21 years old) are more valuable to advertisers than others. As a result, advertisers with low budgets are likely to lose auctions for valuable users than higher budget advertisers. And if there is a correlation between valuable users and protected classes, this can lead to discriminatory ad delivery from the budget size alone (disregarding any possible bias in the algorithm itself). [28]

The researchers also had difficulty in determining whether Facebook's algorithm led to skewed ad delivery since they were unable to have internal access to the algorithm itself, user data, and targeting data. They also had to separate the market effects mentioned above from the optimization effects of the algorithm, distinguish the ad delivery adjustments measured through user feedback from the initial ad classification, and develop techniques to determine the racial breakdown of the delivery audience (since Facebook doesn't provide that information). [28]

In order to determine racial breakdown of the ad audience, the researchers used "Designated Market Area" as a proxy for race by correlating voter records (which usually have racial data and phone numbers that can be linked to Facebook users). [28] Similarly, it is important to note that the study was experimental in nature: the researchers produced their own ads and recorded the results of how those ads did, rather than obtain information from advertisers about their ads in general.

Facebook's Ad Platform

Facebook ads work like many other online advertising platforms: content creators (advertisers) have a budget for each ad and right before a Facebook user is going view an ad, an automated auction occurs to determine which ad is shown. However, Facebook's platform allows advertisers to customize their bids based off user data and demographics (though targeting categories are limited for certain ad categories—housing, employment, credit—after accusations of ad discrimination) as well as define exact users (using phone numbers) to target as well as allow for optimization based of views, clicks, sales, etc. [28], [32]

However, rather than basing the winner for auctions solely on the money spent for the bid, Facebook uses a metric called *total value* which is comprised of the bid value, ad relevance to the user, ad quality, estimated action rates (how often/likely a user will interact with the ad), and other factors. [28], [33] This process is completely closed off to advertisers and is likely done algorithmically given the number of ads displayed on Facebook at a given time.

Major Findings

Skewed delivery of ads were found to occur from budget sizes alone (probably due to market effects as explained prior): ads with very low budgets on average had an audience of over 55% men while ads with high budgets on average had an audience of over 55% women. [28] This suggests that women are considered more valuable when it comes to ad exposure, perhaps due to factors such as click-through rate, etc.

Ads using the same targeting settings but contained content (i.e. images and text) that were stereotypically of interest to men (such as bodybuilding) were delivered to an audience of over 80% men while ads with content that was stereotypically of interest to women (i.e makeup ads) were delivered to an audience of over 90% women. [28] Similar skewed delivery was found about cultural content that was stereotypically more relevant of Black users (i.e. hip-hop music which was shown to an audience of over 85% Black users) while content that was stereotypically of more interest to White users (i.e. country music) was shown to an audience of more than 80% white users. [28]

The researchers noted that the image associated with the ad was found to have the biggest impact on the audience it was delivered to. In fact, even after the researchers made the image transparent to human eyes, there was still a statistically significant skew in ad delivery which suggests that an automated process was the likely culprit for causing the skew and not user interaction itself. [28] Similarly, when

changing the image halfway through the experiment, the delivery changed as well, which suggests that the way Facebook classifies ads is opaque. [28]

It is crucial to note that this type of skewed delivery is partially the purpose of the algorithm itself—delivering ads to a relevant audience—and not inherently discriminatory by itself. In most cases, it shouldn't be considered any different than a traditional advertiser placing ads for a New York City based event in publications based in New York or similar location-based strategies used by online ad algorithms themselves.

However, perhaps most importantly, real-world employment and housing ads can experience significantly skewed delivery. When optimizing for clicks, ads with the same targeting options delivered vastly different racial/gender audiences based off the ad alone.

For example, jobs in the lumber industry reached an audience that was 72% white and 90% male while jobs for janitor positions reached an audience that was 65% female and over 75% black. [28] A similar skew was found for housing ads: an audience of roughly 72% black users were shown a luxury rental ad while an ad for cheap housing to buy had an audience of roughly 49% white users. [28]

Although it is hard to see the correlation the ad delivery algorithm is making between race and housing, it is clear that there is a statistically significant relationship (otherwise the audiences for all types of housing ads would be consistent among races). And although this research cannot be generalized to all ads or even just all employment/housing ads, because the skews are significant, the researchers suggested that the skew most likely occurs with real ads as well. [28] And because this skew occurs with protected classes in categories such as housing and employment ads, there is potential for the algorithm to be encouraging discrimination in the legal sense. [28]

Conclusions

There is a fundamental difference between Facebook and other online advertising and traditional media advertising (such as TV, radio, print, etc.): in traditional advertising, advertisers have the ability to purposefully advertise to a wide and diverse audience and be assured that their ads will reach that audience. [28]

However, this is clearly not the case with Facebook advertising since the ads can be delivered to a skewed audience in unexpected ways that cannot really be prevented. Similarly, the researchers note that an individual's agency to see ads targeted at different groups of people that they do not belong to is severely limited in online

advertising compared to other forms of media. [28]. For instance, a liberal consumer can go watch a conservative television show or buy a conservative publication to view traditional ads that are geared at conservatives. This is virtually impossible to do on Facebook and other online platforms since you cannot change/view your interest profile. As a result, it is also harder to have public scrutiny of online ads because people can only see ads that Facebook decides to deliver to you. [28], [34]

In spite of these difficulties, lawmakers must consider the policy implications on regulating online advertising. For instance, because discrimination can arise independent of the targeting tools that advertisers can use (i.e. through ad delivery algorithms as explained above), lawmakers must think of regulations that do not simply limit these targeting tools. [28]

Similarly, there has been recent debate over whether Facebook should remain protected by Section 230 of the Communications Decency Act. [28], [35], [36] Section 230 states that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another content provider" [37] Although this law has mainly been used by Facebook and other social media platforms to protect them from being held liable for the content of their users, it can be argued that the law also protects Facebook from the actions of its own advertising platform because it cannot be treated as the publisher of the ads themselves. [38] This is another way that Facebook and online advertising is different from traditional advertising: if a newspaper/TV station were found to inadvertently hide ads about employment/housing to protected classes, they would be liable as publishers. As a result, it is important for lawmakers to consider the implications that changes to Section 230 can have on advertising during any talks of reform or revoke the protections on platforms like Facebook.

There should also be greater emphasis on transparency of online advertising platforms in order to allow for greater scrutiny over how ad delivery works and allow users to understand why they are seeing the ads they are. [28]

Finally, it may be possible to remove skewed ad delivery with minimum cost to ad revenue for platforms like Facebook. Researchers at Yale University and IIT Kanpur were able to introduce constraints on the allocations achieved in sub-populations (i.e. white users, black users, male users, female users, etc.) that ensured that an ad was delivered across all sub-populations, allowing for more equal exposure to ad content.[39] These constraints were found to only introduce a minor loss in revenue [39], meaning that online advertising platforms like Facebook could research and formulate methods of preventing potential ad discrimination while retaining their ad revenue.

Healthcare Algorithm

High-risk health care management programs are used across the United States to customize and provide better care for patients in more severe health circumstances. The National Association of Community Health Centers defines high-risk patients as patients "with multiple risk factors that, if left unmanaged" would result in their conditions getting even worse. [40] High risk patients typically make up 20% of the patient population and require one-on-one support in managing healthcare needs. [40]

However, management programs are typically expensive for health care providers so most health care systems rely on algorithms to identify who will benefit from enrollment the most. [41] In order to do so, health systems make a key assumption: those with the greatest care needs will benefit the most from high-risk care management programs.

In this section, we will discuss how researchers were able to find racial bias in an algorithm used by one U.S. health system, despite race not being explicitly used as a factor by the algorithm as well as how such bias can arise from reasonable assumptions.

Science Study

In October 2019, researchers from the University of California, Berkeley, University of Chicago, Brigham and Women's Hospital (Boston), and Massachusetts General Hospital published an article in the journal *Science* where they discovered that an algorithm used by a major U.S. health system had racial bias, assigning the same level of risk to black patients that were sicker than white patients. [41]

The researchers first identified all primary care patients enrolled in risk-based contracts from 2013-2015 in order to form race categories. However because these categories were based on patient self-reporting, the study was unable to look at the impact of the algorithm on intersectional racial and ethnic identities. [41] 6,079 black patients and 43,539 white patients were included as part of the study, with 71.2% enrolled in commercial insurance and 28.8% enrolled in Medicare. On average, patients were 50.9 years old and 63% female. [41]

The researchers wanted to determine whether the algorithm had calibration bias which occurs when the risk score can have different interpretations depending on the racial group. Using the algorithmic risk score for a given patient and the number

of active chronic conditions a patient was experiencing (known as a "comorbidity score") as well as previous years insurance claims, researchers attempted to see how well calibrated the algorithm was for health outcomes and costs. [41]

When using these metrics, it was determined that black patients were more ill than white patients when grouped in the same risk score by the algorithm. Black patients had 26.3% more chronic illnesses than white patients at the highest risk group (97th percentile of risk scores). [41] As a result, black patients that were more sick were placed in lower risk categories and therefore less likely to be enrolled in the risk management programs that could benefit them greatly.

Simulations that accounted for this disparity in the algorithm's risk calculations, the fraction of black patients that were accepted into high-risk health management programs increased from 17.7% to 46.5%. [41] How is this possible when the algorithm excludes racial information? Just like with COMPAS, the algorithm was able to find a proxy for race: total medical expenditures in a year. [41]

At a given level of health (measured by the cormorbidity score), black patients generate lower costs than white patients—roughly $1801 less on average. [41] Similarly, black patients also generate different kinds of costs than white patients, such as fewer in-patient surgical and oupatient specialist costs and more costs related to emergency room visits and dialysis. [41]

As a result, the driving force behind the bias in the algorithm is the disparity between medical costs generated by black patients and white patients. This means that any algorithm which uses an accurate prediction of costs to determine who should be in these health management programs will inherently be racially biased unless they take proper steps to mitigate the bias. [41]

But what causes this disparity between health care costs to begin with? Poorer patients face substantial barriers to accessing health care, even when enrolled in insurance plans such as differential access to transportation to make appointments, competing demands between jobs and childcare, etc. [42] As a result, to the extent that race and socioeconomic status are correlated, these factors affect Black patient uniquely.

Direct discrimination and changes to the doctor-patient relationship also play a part in the cost disparity. A study shows that black patients had an higher uptake of recommended care when the health care provider was black compared to when they were white. [43] Similarly, black patients have reduce trust in the health care system as a result of events such as the Tuskegee study which leads to less patient interaction and therefore lower costs. [41], [44], [45]

The findings of this study emphasize an important part of algorithm design: choosing the right label for the algorithm to be trained on. Although the choice of future costs was reasonable since the goal of the algorithm was to reduce costs, this choice led to the racial bias found by the researchers. Similarly, the top 10 most popular health care algorithms also use cost prediction as their accuracy metric. [41]

Similarly, health in general is holistic and it is hard to use discrete variables (including comorbidity score) to predict one's health. Health care costs (through insurance claims) particularly come from a result of a complex aggregation process with distortions due to structural inequality, incentives, and inefficiencies. [41]

As a result, it is important to note that the algorithm isn't used in isolation: if your risk score was past the 55th percentile, doctors are presented with contextual information to consider enrolling you into the program. The actual enrolled population was 19.2% black (compared to 11.9% of the entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. [41] A "race-blind" enrollment based on algorithmic score alone would yield an enrolling population that is 18.3% black which suggests that although doctors do redress part of the algorithm's bias, they do so far less than a possible algorithm that was trained on a different label (such as avoidable future costs). [41]

Luckily in the case of the algorithm discussed in the study, the manufacturer conducted an independent analysis and found the same bias and after changing the label to integrate health prediction as well as cost prediction, there was an 84% reduction in bias. [41] However, label bias remains pernicious throughout all algorithms as it can be caused from reasonable choices in what the label should be for a given problem.

Conclusion

We conclude by briefly discussing future ways to mitigate biases in artificial intelligence as well as attempts made by governments and organizations in the past including the legal "right of explanation" described by General Data Protection Regulation (GDPR) implemented by the European Union in 2018, ethical principles and guidelines created by private companies and government agencies that use algorithms, and the committee created by the City of New York tasked with coming up with regulations for artificial intelligence used by the City government.

GDPR's "right to explanation"

In 2016, the European Union adopted the General Data Protection Regulation (GDPR) which replaced the 1995 Data Protection Directive passed during the start of the Internet.[46] Included in its multiple provisions is the right for data subjects to receive meaningful information about the logic involved and possible consequences of automated decision-making systems known as the "right to be informed" and the "right not to be subject to automated decision-making." [47]

However, the common interpretation of the "right of explanation" where a person is granted an explanation regarding a specific automated decision made by an algorithm or artificial intelligence agent after the decision is made is not necessarily provided by the GDPR itself. The strongest form of the "right to explanation" is in Recital 71 of the GDPR which is an optional legal interpretation that has no legal binding and although the "right to be informed" is present in the legally binding Article 15 of the GDPR, it was similarly present in the 1995 Data Protection Directive and has proven to not be an effective transparency mechanism. [47]

Researchers at the Oxford Internet Institute recommended that the following steps could be taken in order to rectify deficiencies found in the GDPR [47]:

- Make the right to explanation legally binding rather part of an explanatory Recital.
- Clarify the meaning of "meaning information" that is given to data subjects as well as what the "logic involved" and "significance" of an automated decision means so that it can be practically useful as a regulation against artificial intelligence and biases in the decision-making process.
- Introduce an auditing mechanism to prevent trade secrets from being used as a method to prevent useful explanations from being given out to data subjects.
- Provide support for further research into the feasibility of alternative accountability mechanisms.

Ethical Guidelines

Corporate and public institutional ethical guidelines regarding the use of algorithms and automated decision agents have become more popular in the past decade. An

analysis done by researchers at ETH Zurich found that most corporate and government guidelines focused on transparency, justice and fairness, and freedom and autonomy as the primary values and principles mentioned. [48]

However, many of theses guidelines differ with regards to how to implement such values into actual algorithms. For instance, there is great variance amongst guidelines regarding what information should be disclosed or communicated to users (i.e. whether it should be the source code itself, the uses of the AI, the usage of personal data, etc.). [48]

As a result, it is impossible to find a general consensus among corporate and public guidelines to be used as a regulation overall for artificial intelligence. Other attempts at creating guidelines with regards to managing algorithms are the FAIR principles established by academia which emphasize transparency, interoperability, and reusability for both humans and machines. [49]

These principles were found to be better than transparency alone (i.e. forcing algorithms in academia to be open-sourced) as usually those schemes do not have a relevant audience that can understand them and can allow for exploits to be used against academic research. Furthermore, there should be a greater emphasis on empirical analysis regarding different transparency since more information doesn't always yield better results unless the audience is properly informed. [49]

New York City Automated Decision Systems (ADS) Task Force

On May 18th, 2018, Mayor Bill de Blasio announced that the New York City would create a task force dedicated to exploring how the City utilizes "automated decision systems" (i.e. algorithms) and determine guidelines to promote transparency, fairness, and equity. [50]

The announcement was notable due to it being the first taskforce in the United States created by a city government for reviewing the use of algorithms as well as the scope and scale of task force's mission. The task force consisted of members from the Mayor's Office and various academics from the fields of technology and legal experts and could have served as an example of how governments could elegantly determine guidelines for the use of algorithms by their agencies. [50]

However, the committee was plagued with multiple issues: primarily misunderstanding on the scope that possible regulations (i.e is an Excel script an automated decision system?) as well City officials preventing information about what automated

systems were currently used by the government from being disclosed to the taskforce, preventing practical regulations that could have a noticeable effect. [51]

Governments are the biggest customers for algorithmic systems and as a result, critics state that forcing a transparency standard could just force companies to not work on government contracts. [52] But on the other hand, that results in governments themselves having the most leverage with regards to transparency and a possible solution is to include transparency clauses in government contracts with private systems companies. [52]

Final Thoughts

It is clearly apparent that algorithmic systems and artificial intelligence contain biases caused by multiple factors such as the inability to frame a goal into a quantifiable model that can be used by artificial agents and have real world implications such as COMPAS algorithm recommending harsher sentences for Black defendants.

These issues are likely to be exacerbated as more processes around the world become automated through algorithmic systems and artificial intelligence. Although studies show that companies respond to public naming of algorithmic biases and disparities and having a more holistic approach when creating algorithmic systems (such as including more variables to represent more factors), it is clear that more research should be placed into determine empirical guidelines on how artificial intelligence systems should be regulated to reduce biases as well as educating the public on the imperfections that artificial intelligence systems have. [53], [54]

References

[1] I. D. Corporation, *Idc forecasts improved growth for global ai market in 2021*, `https://www.idc.com/getdoc.jsp?containerId=prUS47482321`, Feb. 2021.

[2] B. M.
bibinitperiod Co., *Are alexa and siri considered ai?* `https://bernardmarr.com/default.asp?contentID=1830`.

[3] M. Panzarino, *Apple combines machine learning and siri teams under giannandrea*, `https://techcrunch.com/2018/07/10/apple-combines-machine-learning-and-siri-teams-under-giannandrea/`, Jul. 2018.

[4] Tesla, *Autopilot*, `https://www.tesla.com/autopilotAI`.

[5] G. Trends, *Worldwide trend of "deep learning" from 4/26/11 to 4/26/2021*, https://trends.google.com/trends/explore?date=2011-04-26\%202021-04-26\&q=deep\%20learning.

[6] ——, *Worldwide trend of "neural net" from 4/26/11 to 4/26/2021*, https://trends.google.com/trends/explore?date=2011-04-26\%202021-04-26&q=neural\%20net.

[7] ——, *Worldwide trend of "reinforcement learning" from 4/26/11 to 4/26/2021*, https://trends.google.com/trends/explore?date=2011-04-26\%202021-04-26&q=reinforcement\%20learning.

[8] P. Wang, "On defining artificial intelligence," *Journal of Artificial General Intelligence*, vol. 10, no. 2, 1–37, 2019. DOI: doi:10.2478/jagi-2019-0002. [Online]. Available: https://doi.org/10.2478/jagi-2019-0002.

[9] I. C. Education, *What is machine learning?* https://www.ibm.com/cloud/learn/machine-learning, Jul. 2020.

[10] S. F. Deangelis, *Artificial intelligence: How algorithms make systems smart*, https://www.wired.com/insights/2014/09/artificial-intelligence-algorithms-2/.

[11] A. Howard, "Are we trusting ai too much? examining human-robot interactions in the real world," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20, Cambridge, United Kingdom: Association for Computing Machinery, 2020, 1, ISBN: 9781450367462. DOI: 10.1145/3319502.3374842. [Online]. Available: https://doi.org/10.1145/3319502.3374842.

[12] K. Hammond, *5 unexpected sources of bias in artificial intelligence*, https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/, Dec. 2016.

[13] B. X. Chen, *Hp investigates claims of 'racist' computers*, https://www.wired.com/2009/12/hp-notebooks-racist/, Dec. 2009.

[14] A. Technica, *Tay, the neo-nazi millennial chatbot, gets autopsied*, https://arstechnica.com/information-technology/2016/03/tay-the-neo-nazi-millennial-chatbot-gets-autopsied/, Mar. 2016.

[15] S. Jeong, *How to make a bot that isn't racist*, https://www.vice.com/en/article/mg7g3y/how-to-make-a-not-racist-bot, Mar. 2016.

[16] C. Sinders, *Microsoft's tay is an example of bad design*, https://medium.com/@carolinesinders/microsoft-s-tay-is-an-example-of-bad-design-d4e65bb2569f, Mar. 2016.

[17] S. Flaxman, S. Goel, and J. M. Rao, "Filter bubbles, echo chambers, and online news consumption," *Public Opinion Quarterly*, vol. 80, no. S1, 298–320, 2016. DOI: 10.1093/poq/nfw006. [Online]. Available: https://doi.org/10.1093/poq/nfw006.

[18] K. Hao, *This is how ai bias really happens—and why it's so hard to fix*, https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/, Feb. 2019.

[19] B. E. Harcourt, "Risk as a proxy for race," *University of Chicago Law & Economics Olin Working Paper*, Sep. 2010.

[20] E. P. I. Center, *Algorithms in the criminal justice system: Risk assessment tools*, https://epic.org/algorithmic-transparency/crim-justice/.

[21] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine bias*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

[22] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, *How we analyzied the compas recidivism algorithm*, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, May 2016.

[23] U. S. C. Bureau, *Annual estimates of the resident population: April 1, 2010 to July 1, 2019*, https://data.census.gov/cedsci/table?q=Broward\%20County,\%20Florida\&tid=PEPPOP2019.PEPANNRES, Census Bureau Annual Estimate for Broward County.

[24] U. S. S. Commission, *The Effect of Aging on Recidivism Among Federal Offenders*, https://www.ussc.gov/research/research-reports/effects-aging-recidivism-among-federal-offenders, Dec. 2017.

[25] A. R. Piquero and R. W. Brame, "Assessing the Race–Crime and Ethnicity–Crime Relationship in a Sample of Serious Adolescent Delinquents," *Crime & Delinquency*, vol. 54, no. Issue 3, Apr. 2008.

[26] H. L. Review, "State v. Loomis," *Harvard Law Review*, vol. 130, no. 5, Mar. 2017.

[27] M. Charts, *US Advertising Media Market Sizes ($B), 2020 v. 2024*, https://www.marketingcharts.com/advertising-trends-114887.

[28] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes," *CoRR*, vol. abs/1904.02095, 2019. arXiv: 1904.02095. [Online]. Available: http://arxiv.org/abs/1904.02095.

[29] R. Iyengar, *Here's how big Facebook's ad business really is*, https://www.cnn.com/2020/06/30/tech/facebook-ad-business-boycott/index.html, Jul. 2020.

[30] K. Benner, G. Thrush, and M. Isaac, *Facebook engages in housing discrimination with its ad practices, u.s. says*, https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html, Mar. 2019.

[31] Facebook, *About ad relevance diagnostics*, https://www.facebook.com/business/help/403110480493160?id=561906377587030.

[32] ——, *Choose the right objective*, https://www.facebook.com/business/help/1438417719786914.

[33] ——, *About ad auctions*, https://www.facebook.com/business/help/430291176997542?id=561906377587030.

[34] Mozilla, *Facebook's ad archive api is inadequate*, https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate/, Apr. 2019.

[35] *Section 230: A Key Legal Shield For Facebook, Google Is About To Change*, https://www.npr.org/sections/alltechconsidered/2018/03/21/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change, Mar. 2018.

[36] J. Guynn, *Donald Trump and Joe Biden vs. Facebook and Twitter: Why Section 230 could get repealed in 2021*, https://www.usatoday.com/story/tech/2021/01/04/trump-biden-pelosi-section-230-repeal-facebook-twitter-google/4132529001/, Jan. 2021.

[37] E. F. Foundation, *Section 230 of the Communications Decency Act*, https://www.eff.org/issues/cda230.

[38] R. T. Ring, J. H. Blavin, J. Patashnik, and E. A. Kim, *Defendant's notice of motion and motion to dismiss first amended complaint; memorandum of points and authorities in support thereof*, https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf, Jun. 2017.

[39] L. E. Celis, A. Mehrotra, and N. K. Vishnoi, *Toward controlling discrimination in online ad auctions*, 2019. arXiv: 1901.10450 [cs.GT].

[40] N. A. of Community Health Centers, *Value transformation framework action guide*, https://www.nachc.org/wp-content/uploads/2019/03/Risk-Stratification-Action-Guide-Mar-2019.pdf, Jul. 2019.

[41] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, 447–453, 2019, ISSN: 0036-8075. DOI: 10.1126/science.aax2342. eprint: https://science.sciencemag.org/content/366/6464/447.full.pdf. [Online]. Available: https://science.sciencemag.org/content/366/6464/447.

[42] K. Fiscella, P. Franks, M. R. Gold, and C. M. Clancy, "Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care," *JAMA*, vol. 283, no. 19, 2579–2584, May 2000, ISSN: 0098-7484. DOI: 10.1001/jama.283.19.2579. eprint: https://jamanetwork.com/journals/jama/articlepdf/192714/jpp90039.pdf. [Online]. Available: https://doi.org/10.1001/jama.283.19.2579.

[43] M. Alsan, O. Garrick, and G. C. Graziani, "Does diversity matter for health? experimental evidence from oakland," National Bureau of Economic Research, Working Paper 24787, Jun. 2018. DOI: 10.3386/w24787. [Online]. Available: http://www.nber.org/papers/w24787.

[44] M. Alsan and M. Wanamaker, "Tuskegee and the health of black men*," *The Quarterly Journal of Economics*, vol. 133, no. 1, 407–455, Aug. 2017, ISSN: 0033-5533. DOI: 10.1093/qje/qjx029. eprint: https://academic.oup.com/qje/article-pdf/133/1/407/30636482/qjx029.pdf. [Online]. Available: https://doi.org/10.1093/qje/qjx029.

[45] K. Armstrong, K. L. Ravenell, S. McMurphy, and M. Putt, "Racial/ethnic differences in physician distrust in the united states," *American Journal of Public Health*, vol. 97, no. 7, 1283–1289, 2007, PMID: 17538069. DOI: 10.2105/AJPH.2005.080762. eprint: https://doi.org/10.2105/AJPH.2005.080762. [Online]. Available: https://doi.org/10.2105/AJPH.2005.080762.

[46] E. D. P. Supervisor, *The history of the general data protection regulation*, https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en.

[47] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *SSRN Electronic Journal*, 2016. DOI: 10.2139/ssrn.2903469. [Online]. Available: https://doi.org/10.2139/ssrn.2903469.

[48] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, 389–399, 2019.

[49] J. Kemper and D. Kolkman, "Transparent to whom? no algorithmic accountability without a critical audience," *Information, Communication & Society*, vol. 22, no. 14, 2081–2096, 2019. DOI: 10.1080/1369118X.2018.1477967. eprint: https://doi.org/10.1080/1369118X.2018.1477967. [Online]. Available: https://doi.org/10.1080/1369118X.2018.1477967.

[50] C. of New York, *Mayor de blasio announces first-in-nation task force to examine automated decision systems used by the city*, https://www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by, May 2018.

[51] A. F. Cahn, *The first effort to regulate ai was a spectacular failure*, https://www.fastcompany.com/90436012/the-first-effort-to-regulate-ai-was-a-spectacular-failure, Nov. 2019.

[52] J. Powles, *New york city's bold, flawed attempt to make algorithms accountable*, https://www.newyorker.com/tech/annals-of-technology/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable, Dec. 2017.

[53] I. D. Raji and J. Buolamwini, "Actionable auditing," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Jan. 2019. DOI: 10.1145/3306618.3314244. [Online]. Available: https://doi.org/10.1145/3306618.3314244.

[54] I. Chen, F. D. Johansson, and D. Sontag, *Why is my classifier discriminatory?* 2018. arXiv: 1805.12002 [stat.ML].