manuscript No.
(will be inserted by the editor)

Algorithms are not Neutral

Bias in Collaborative Filtering

Catherine Stinson

Received: date / Accepted: date

Abstract When Artificial Intelligence (AI) is applied in decision-making that affects people's lives, it is now well established that the outcomes can be biased or discriminatory. The question of whether algorithms themselves can be among the sources of bias has been the subject of recent debate among Artificial Intelligence researchers, and scholars who study the social impact of technology. There has been a tendency to focus on examples where the dataset used to train the AI is biased, and denial on the part of some researchers that algorithms can also be biased. Here we illustrate the point that algorithms themselves can be the source of bias with the example of collaborative filtering algorithms for recommendation and search. These algorithms are known to suffer from cold-start, popularity, and homogenizing biases, among others. While these are typically described as statistical biases rather than biases of moral import; in this paper we show that these statistical biases can lead directly to discriminatory outcomes. The intuitive idea is that data points on the margins of distributions of human data tend to correspond to marginalized people. The statistical biases described here have the effect of further marginalizing the already marginal. Biased algorithms for applications like media recommendations can have significant impact on individuals' and communities' access to information and culturally-relevant resources. This source of bias warrants serious attention given the ubiquity of algorithmic decision-making.

Philosophy Department
Queen's University
312 Watson Hall
Kingston, ON K7L 3N6, Canada
E-mail: c.stinson@queensu.ca

School of Computing
Queen's University
557 Goodwin Hall
Kingston, ON K7L 2N8, Canada

Keywords algorithms · machine learning · statistical bias · discrimination · recommendation · search

1 Introduction

There is growing awareness that the outcomes of algorithmic processes can be discriminatory. The best known recent examples of algorithmic discrimination happen to be ones where the data used to train machine learning algorithms are systematically biased, leading to algorithms with discriminatory outcomes. Several cases have been uncovered where using data about past decisions to train systems to make policing (Angwin et al, 2016; Cardoso, 2020), hiring (Ajunwa et al, 2016; Raub, 2018), medical (Ferryman and Pitcan, 2018; Obermeyer et al, 2019), or other decisions in the present means that historical discrimination gets baked into the algorithm, perpetuating the bias in the next generation of decisions.

Many of the suggested approaches for mitigating algorithmic bias involve de-biasing datasets. For instance, Ajunwa et al (2016) outline data modification processes that can prevent discriminatory decisions in the context of hiring, and Ferryman and Pitcan (2018) suggest ways of diversifying medical datasets in order to prevent bias. Obermeyer et al (2019) are an exception to this pattern; they outline an alteration to the algorithm used to score patients' health needs to fix the underestimation of Black patients' illness severity. Sánchez-Monedero et al (2020) review and evaluate several methods used to mitigate bias in hiring algorithms. The general literature on algorithmic fairness tends to remain agnostic as to the root causes of unfairness in algorithms. For a critical review of this literature, see Mitchell et al (2021). The main concern here is where in the workflow from data collection and algorithm design to testing and implementation the causes of discriminatory outcomes can be found. Because those causes are not restricted to the data preparation phases, de-biasing datasets is not always a viable strategy for mitigating algorithmic bias.

One complication in pinpointing exactly how and where algorithms are biased is the fact that bias has several different meanings. Some of these are value-neutral, like technical definitions of bias in statistics, while others have a moral character, implying either purposeful or unconscious discrimination. Both broad types of bias will be implicated in demonstrating that algorithms themselves can be biased. That algorithms introduce statistical biases is relatively uncontroversial. One of the main claims to be defended is that statistical bias affecting algorithms can cause discriminatory outcomes.

In Section 2, we examine a triad of options as to where bias might be located: data, people, and algorithms. In Section 3 we introduce collaborative filtering, and outline some of the most well known statistical biases known to affect this class of recommendation algorithms, including the cold-start problem, popularity bias, over-specialization, and homogenization. In Section 4 we offer evidence for a selection bias affecting iterative information filtering al-

gorithms generally. Section 5 connects the statistical biases so far outlined to empirical evidence suggesting that these algorithms produce biased outcomes in recommendation and search systems. Section 6 offers some concluding thoughts on why recognizing bias in algorithms themselves is important.

2 Data, People, or Algorithms?

That algorithms themselves are neutral is a popular refrain among AI researchers. In an interview, deep learning pioneer Yoshua Bengio insisted that "The algorithms we use are neutral" (Groen, 2018). On Twitter, Yann LeCun declared, "People are biased. Data is biased... But learning algorithms themselves are not biased" (LeCun, 2019), then later doubled down on the claim, tweeting, "ML systems are biased when data is biased..." in response to a controversy over a photo upsampling program that seemed to systematically render blurry people of color (POC) as white (LeCun, 2020).

2.1 Biased Data

The examples we began with are ones where historical discrimination in domains like policing, hiring, and health care led to biased datasets, which when used to train a machine learning classifier, automated and reproduced those historical wrongs in another generation. In addition to biased datasets that result in this way from systemic discrimination on the level of societies, biased datasets can also be the downstream result of a different kind of systemic discrimination. Because of a lack of gender and racial diversity (among other axes of difference) in AI as a field, developer teams often lack diversity. That facial recognition algorithms are an order of magnitude less accurate for Black female faces than for white male faces has been attributed to the lack of Black and female faces among the training examples used to build facial recognition systems. That lack of diversity in the training examples is in turn thought to stem from a lack of gender and racial diversity among AI researchers (Buolamwini and Gebru, 2018), either because computer vision datasets tend to start with pictures of lab members, because developers looking for data tend to look in places where they themselves might post pictures, or because media representations are less diverse than the general population.

In these cases of biased datasets, 'bias' can have several distinct meanings. In statistics and machine learning, 'selection bias' refers to a non-random process being used to select a sample from a population. That the datasets used to train facial recognition software oversample white male faces compared to the population the software will be used on is a selection bias. A selection bias in the data sampling often leads to poorer performance of models built with those data (even when that selection bias has no moral implications).

The colloquial meaning of bias is closer to the definition Friedman and Nissenbaum offer of "bias of moral import," which is, "systematically and

unfairly discriminat[ing] against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996). In the same example, the fact that false identifications of faces are significantly higher for Black people than white leads to unfair discrimination when facial recognition software is used for purposes like finding crime suspects in crowds. Proportionally more innocent Black people will be stopped and risk being falsely arrested than for other groups. There have already been several documented cases of Black men being falsely arrested due to inaccurate facial recognition.

As the facial recognition example shows, these two kinds of bias (statistical and moral) can interact. Here is another example: If a police force stops and frisks Black men without cause more often than other people (which has been demonstrated to be the case in several jurisdictions in North America and Europe), that would lead to proportionally more charges for petty crimes among that demographic group (or proportionally fewer charges for petty crimes for other groups). This would be unfair discrimination. In this example too, a selection bias is present, since the people being stopped are not chosen at random, and because of the selection bias, discriminatory harm occurs.

2.2 Biased People

Another potential source of algorithmic bias is the people building algorithms. There are documented cases where algorithms have been designed specifically to create discriminatory outcomes. Redlining certain neighbourhoods as high risks for mortgages, based on the racial composition of residents, gerrymandering election districts to disenfranchise some types of voters, or choosing to target only men to show certain kinds of job ads (Dwoskin, 2018) are three examples.

In most cases biased algorithm builders are presumably not motivated by overt discrimination. A more common scenario is that products are built for the benefit of one group, while inadvertently producing negative side-effects for others, such as how YouTube's click-maximizing algorithm benefits advertizers at the expense of website users (Tufekci, 2018), or how online proctoring software is built to meet the needs of university administrators at the expense of students (Cahn et al, 2020).

Quite often bias is accidental and unforeseen, resulting from the limited perspective of algorithm makers and business owners rather than gross negligence. Speech recognition algorithms that fail to work for people with non-standard accents or difficulty speaking are one example. The choice of research questions to pursue, or applications to develop can also overlook the needs of people not on the radar of algorithm makers and business owners. An example is how Apple's health app was initially released without including a period tracker, despite that being one of the primary uses for a health app among people with uteruses (Perez, 2015). The oversight processes in place to ensure safety, usefulness and performance of an algorithm can likewise fail to consider the needs of some groups, like automatic soap dispensers that fail to reliably

detect dark hands (Fussell, 2017). Even when algorithm builders think they are aiming for inclusion, the needs of minority users and the systems of oppression in operation might not be well understood by those developing the algorithms.

2.3 Biased Algorithms

Algorithms themselves are a source of bias that is often overlooked in reviews of algorithmic bias (Barocas and Selbst, 2016; Kirkpatrick, 2016), which tend to focus on discriminatory outcomes or data bias. Some articles mention the possibility of bias in algorithms themselves, but only offer examples of biased data or biased people (Garcia, 2016). An exception is Danks and London (2017), who include "algorithmic processing bias" in their taxonomy of kinds of bias. Their focus is bias in autonomous systems like self-driving cars, and offer the example of using a biased estimator in order to minimize variance if you have a small sample size. Another exception is Hooker (2021), who makes a plea for looking beyond just data bias, arguing that model design also contributes to bias, citing examples from work in facial recognition.

Here we expand on the point that algorithms themselves can be biased, and apply it to the context of a popular class of algorithms used in recommendation and search tasks. Previous studies of algorithmic bias in recommender systems follow the general pattern of treating algorithmic bias as a matter of outcomes, or as a data problem (Edizel et al, 2020). Below we review the statistical biases known to exist in collaborative filtering algorithms, as well as a selection bias inherent in a more general class of algorithms they fall into, suggesting that the phenomenon is quite widespread. We then offer examples of how those statistical biases translate into bias of moral import, particularly for marginalized users of recommender systems and other information filtering systems like search engines.

3 Bias in Collaborative Filtering

Collaborative filtering algorithms are used in popular recommender systems like Amazon and Netflix, that show users items based on criteria like "Customers who viewed this item also viewed" or "Because you watched..." To generate these recommendations, first user profiles are constructed based on a person's explicit ratings of media or products, such as likes or stars, as well as their implicit ratings generated from activity like clicks or viewing time. To find recommendations suitable for a user with that set of likes and dislikes, their profile is compared to other users' profiles to find close matches. Items that were rated highly by other users with similar profiles, but that have not been seen by the current user, are then recommended to that user. User profiles are models of user preferences, and are regularly updated as the user interacts with the system, with the goal of making the profile a more accurate predictor of the user's behaviour over time.

Collaborative filtering can be contrasted with content-based recommendation algorithms which might instead look for similarities between the content a user likes and other available content. Collaborative filtering depends on the assumption that no user is unique, in that recommendations happen through matching with other users. Where that assumption is violated (users who don't share the same tastes with anyone else in the system), collaborative filtering can be expected to work poorly. If the most unique users turn out to be people who belong to multiple minority groups, so there are a priori reasons for expecting that collaborative filtering might be biased in favour of the majority. Below we outline a few of the specific ways in which collaborative filtering algorithms are known to be biased. Olteanu et al (2019) catalogue a number of additional biases that can occur at all stages of the software development cycle for recommendation systems.

3.1 Cold-Start Problem

The cold-start problem is perhaps the best known bias affecting collaborative filtering. Ironically, although collaborative filtering was intended as a replacement for human reviewers, recommending new releases is a task collaborative filters are uniquely unqualified to do. When a new item becomes available, there are initially no ratings of it by any user. If there are no ratings of an item by any user, then a collaborative filtering algorithm cannot recommend the item to anyone, since recommendations are based on what other users have rated. The tendency of platforms like Netflix and Amazon to push their new offerings to the top of the recommendation list is somewhat justified, because otherwise they would remain unknown. The job of the critic has been largely replaced by recommendation algorithms, despite these algorithms' inability to do what critics do. Writ large, items that have been in the system longer will build up more ratings over time, so be more likely to be recommended than newer items.

This dynamic where older items are preferentially recommended over newer ones would develop no matter how initial ratings are distributed initially, as long as new material is added over time. This is a case where adding more data to correct for an imbalance in the dataset would be difficult to implement. An obvious approach would be to add synthetic ratings to ensure that all items have a uniform number of ratings. When an item is new, there is little basis on which to create synthetic ratings (without doing content-based recommendation instead of collaborative filtering), so adding synthetic data to bump up items with few ratings could lead to low quality recommendations. Using a hybrid of content-based and collaborative filtering is explored by Schein et al (2002).

From the perspective of users, the cold-start problem appears as a (small c) conservative bias, where popular but older items are hard to avoid, and new things are harder to find. Likewise, the earlier an individual user gives a positive rating to an item, the more of an effect that item will have on

their future recommendations. A youthful preference for *Lady and the Tramp* would affect the user's recommendations, and therefore the user's viewing habits and ratings, for the entire life of their profile, possibly leading to a recommendation for *Dumbo* twenty years later, despite their tastes having matured. In contrast, a more recent interest in documentaries would have fewer total ratings associated with it, and thus exert relatively less of an effect on the user's recommendations. Weighting ratings by recency is a way of mitigating that effect (**?**).

3.2 Popularity Bias

A closely related problem is known as popularity bias (Herlocker et al, 2004; Steck, 2011), where very popular items are likely to get recommended to every user (and since recommendations make ratings more likely, popular items tend to increase in popularity). So even a user whose only positive ratings are for medieval Persian editions of ancient medical texts might get recommendations for *The Very Hungry Caterpillar*, simply because no matter what you buy, it's likely that someone who bought the same has also bought *The Very Hungry Caterpillar*. Relatedly, a user might have bought *Fifty Shades of Gray* because they are writing a dissertation about representations of kink in popular culture, and end up having to wade through pulp romance novel recommendations that come highly rated by *Fifty Shades of Gray* fans, despite having no interest in the genre.

Abdollahpouri et al (2019) show that popularity bias affects some groups of users more than others, with users who prefer mostly "long-tail" items (items that are not popular overall) being most adversely affected. One approach used by **?** to mitigate popularity bias is to add weights to the recommendations, such that when users are more similar, their recommendations are given more weight, and when users are less similar, their recommendations are given less weight. Another approach that may benefit users who prefer long-tail items would be to track the average popularity of a user's highly rated items, and weight the recommendations of items based on their popularity accordingly.

Profile injection attacks manipulate the probability of an item being recommended through the creation of fake user ratings. An infamous example is how the Amazon page for a book by anti-gay televangelist, Pat Robertson, listed an anal sex guide as a recommendation, after pranksters repeatedly viewed the two items together in order to form an association (Olsen, 2002). This trick has also been used as a marketing ploy. Profile injection attacks illustrate the extent to which recommendations depend on popular patterns of ratings of other users.

3.3 Over-specialization

Over-specialization occurs when a recommender algorithm offers choices that are much more narrow than the full range of what the user would like. In

statistical terms this is not a problem of bias but of variance (the expectation of how far a variable deviates from its mean). Adamopoulos and Tuzhilin (2014) treat over-specialization as a problem stemming from an exclusive focus on prediction accuracy, while overlooking user-satisfaction, which might depend also on there being enough variety in recommendations.

Intuitively, the problem arises because items similar to those previously liked by a user will have a high probability of also being liked, even though what the user wants might be a wider range of recommendations that cover their preferences more fully. For example, the user may not want to get stuck in a rut of only watching teen comedies after one nostalgic viewing of *Mean Girls*, even if they do also like *Clueless*, and *Election*.

(Steck, 2011) mitigates this bias by preferentially using items from the tail of a user's rating distribution as the basis for matching profiles. Adamopoulos and Tuzhilin (2014) mitigate both over-specalization and the popularity bias, using a "probabilistic nearest neighbors" method. This involves sampling neighbors probabilistically, weighted based on their distance. This results in user recommendations coming from a variety of distances, which diversifies the recommendations, while still treating closer neighbors as most trustworthy. This method outperforms standard methods on both prediction accuracy and utility-based ranking (which takes into account users' perceptions of the quality of the recommendations.

3.4 Homogenization

Another issue for which there is some scattered evidence is homogenization. Popularity bias refers to how single items that are very popular are over-recommended. Homogenization is an effect over the dataset as a whole, where the variance of items recommended to all users combined decreases over time. One hypothesis for how this may come about is that users' preferences either for diversity or popularity in their media consumption is not captured by collaborative filtering algorithms, as described by Abdollahpouri et al (2019). All users are treated as though they prefer popular media.

A 2008 study found that since online journals became common, which increased the availability of academic literature, citation practices have narrowed. Fewer journals, and fewer articles are being cited, suggesting that people are reading less widely, not more (Evans, 2008). Evans attributes the effect to the greater efficiency of finding sources online, by following a few links, compared to browsing library stacks, where it takes longer to find specific sources, but you end up seeing a greater variety of papers in passing.

A recent study (West, 2019) suggests that GoogleScholar's recommendations may have had a homogenizing effect on citation practices. More citations are going to the top 5% of papers by citation count, and a smaller proportion of papers are being cited overall since the release of GoogleScholar. When the recommendation systems we use are designed to only show us items that

other users have interacted with, rather than sampling from the entire dataset equally, this narrowing of recommendations is likely to happen.

The phenomenon of "filter bubbles" or "echo chambers" is often blamed on the laziness or closed-mindedness of individuals, who can't be bothered to look beyond their social media feeds, or who don't want to do the work of consuming media that might challenge their comfortable opinions. However filter bubbles may arise in part as a result of the homogenization that is characteristic of recommender algorithms. It may not be that users fail to venture outside their bubbles, but rather that the algorithm traps users inside. A comparison of several recommendation algorithms in terms of how author gender affects book recommendations, found that some algorithms produce recommendation lists that are "more imbalanced than the item universe" even when user ratings are more balanced (Ekstrand et al, 2018b). In this case it is abundantly clear that the data is not the only problem. The algorithms is contributing bias over and above any bias that may exist in the data.

4 Selection Bias in Information Filtering

For many popular recommender systems, ratings are sparse relative to the number of items available. For example, most of the items available for sale on Amazon will never be bought or viewed by most users. ML algorithms, including those used to predict user preferences, are very often designed to have high prediction accuracy, which in this case is a measure of the probability that the user will indeed like an item if that item is recommended to the user. It is possible to maximize prediction accuracy without capturing the full variety of items that the user would like, however. Accuracy and precision can trade off. An example might be taking the safe bet of only recommending sequels to a user's favorite movie, and not bothering to try to recommend anything else, which might generate only likes, but would miss many other movies that the user would also like.

If the missing data (i.e., the items that are not rated at all) were missing at random, then the possibility of low precision while accuracy is high would be a minor worry, because the algorithm would not be able to confine its recommendations to just one small corner of the space of possibilities. The ratings are not missing at random in models that are learned iteratively over time from user ratings, however (Stinson, 2002; Chawla and Karakoulas, 2005). As the recommender narrows in on the user's tastes, it is simultaneously narrowing the space of possibilities that it can recommend, and thus the scope of the data available to it on which to improve its model of the user.

Many of the ratings the system gets, whether explicit or implicit, are for items that the user has seen because the system recommended the items. The system cannot learn from the user's hypothetical ratings of things the user has not been shown. In order to do its job well, the algorithm needs a broader base of ratings, including confirmations that the user indeed does not like items that the model predicts the user would not like. In virtue of having its

source of training data tied to the outputs of the user profile it is building, the collaborative filtering system imposes a selection bias on its own training data, then iteratively exacerbates that bias as it improves its model of the user over time.

Several of these biases stem from the number and timing of ratings not being evenly distributed among items in the dataset, and are exacerbated by the fact that recommendations influence what gets seen and therefore rated at later times. These biases affect different users differently, and additional biases originate from the fact that users are not uniformly distributed in preference space. How much the user values novelty, how much the user's tastes have changed from their starting point, and how far their tastes lie from the mean can all vary. Many users' preferences will cluster around popular items, but other users will cluster in smaller niche groups (Horror fans, perhaps), and still others will have rare preferences (like our medieval Persian medical text fan), or atypical combinations of preferences (a fan of both Death Metal and musicals, for example). Neophytou et al (forthcoming) show that the popularity of the items a user likes affects the accuracy of recommendation prediction, such that users with niche tastes get less accurate recommendations.

Collaborative filtering algorithms belong to the broader class of information filtering algorithms. Information filters choose items from information streams to deliver to users based on a model of the user's preferences, or a particular topic. Some common examples are a search engine returning documents that include a user provided search term, or a personalized newsfeed delivering articles on a given topic to a user's inbox. Spam filters are also information filters, but where the selected items are redirected away from users.

Information filters that continuously update their predictive model based on feedback (e.g., what the user clicks on), to improve performance during operation are alternatively called "online," "active," or "iterative". Here we use the term iterative information filtering. The sequence of events is a loop starting with a recommendation step based on the initial model, then the user is presented with the recommendations, and chooses some items to interact with. These interactions provide explicit or implicit feedback in the form of labels, which are used to update the model. Then the loop repeats with recommendations based on the updated model.

The user's interactions change the model, based on what was recommended, which in turn affects what can be recommended at later stages. Just as in the special case of collaborative filtering, iterative information filtering introduces a selection bias (Stinson, 2002; Chawla and Karakoulas, 2005). Since labels are only provided for items that were recommended, the missing at random assumption is violated. This bias is investigated in Sun et al (2018), who refer to it as "iterated algorithmic bias". One of the main effects of the selection bias is more homogeneous recommendations (Sun et al, 2018), narrowing the space of items available for recommendation.

The homogenizing bias occurs in iterative information filtering contexts generally. For some information filtering tasks, it may not be a bad thing for recommendations to become more homogenous over time. If the purpose of

the filter is to find articles relevant to a very particular interest, then it might be desirable for the filter to become progressively better at picking out that one specific topic. But in contexts like GoogleScholar searches, increasingly homogenous search results for a given search term would typically be a negative outcome. For instance, if the user is doing a literature search, they want the full complement of relevant articles, not just the most cited ones. Likewise, if the user is looking for citation information for a specific article, an exact match is more desirable than the most cited match in that neighborhood.

A number of ad hoc strategies are described for mitigating this bias. These include diversifying the collection of items that are used to learn the next iteration of the model by estimating labels for items that were not recommended (Stinson, 2002), and explicitly modeling the censoring mechanism to correct the bias (Chawla and Karakoulas, 2005).

5 Statistical Bias Can Cause Discrimination

Uncorrected statistical bias has negative effects on the performance of algorithms, which is bad for users, as well as media producers and advertizers who stand to gain from accurate recommendations. The negative effects are worse for some users than others, and the implications go well beyond occasionally having to scroll past unwanted recommendations.

As algorithms mediate more and more of our access to information, access to services, and decisions about our lives, their uneven performance can become a significant equity issue. The biases described here have the greatest negative effects on users located at the margins of preference distributions: people with unusual tastes, or unique combinations of tastes. The people on the margins of distributions are literally marginalized people, whom non-discrimination law is supposed to protect (Treviranus, 2014).

People from minority communities have noted that recommender algorithms do not work well for them. Noble (2018) documents the ways that search algorithms fail to serve the needs of black women. One of her examples is a hair salon owner who struggled to get her business to show up as a recommendation on Yelp when you search for "'African American,' 'Black,' 'relaxer,' 'natural,"' as keywords. Complaints about culturally inappropriate recommendations, like white hairdressers being recommended for those search terms, or Christmas movies being recommended to non-Christians, are common online. Popularity and homogenizing biases may be at fault in those examples. A related issue arises when the recommender system does figure out that a user belongs to a minority group, but overfits to an essentialized version of that identity. That you cannot escape ads for *Rupaul's Drag Race* if your online presence reveals any interest in LGBTQ+ issues stems from over-specialization.

There is some empirical evidence for differential effects of algorithmic bias on demographic groups. Mehrotra et al (2017) investigate whether search engines "systematically underserve some groups of users." Ekstrand et al (2018a)

find significant differences in the utility of recommendation systems for users of different demographic groups (binary gender, and age), although not exclusively benefitting the larger groups. Zafar et al (2017) discuss "disparate mistreatment," which arises when a classifier's misclassification rates differ across social groups. An example (which stems from data bias) is how the COMPAS algorithm made more false positive errors with black defendants, labeling people who would not reoffend as being high risk, while making more false negative errors with white defendants (Angwin et al, 2016).

6 Conclusions

Perhaps the greatest source of harm is that the illusion of neutrality algorithms have can be exploited in attempts to roll back protections against discrimination. Appeals to the neutrality of algorithms as a cover for discriminatory outcomes has become a fairly common trope. In 2019 the UK education secretary came under fire for apparent discrimination in the algorithmically calculated A-levels scores that were to replace university entrance exams cancelled because of COVID-19 (Meadway, 2020). Initial government responses to the controversy pointed to "the algorithm" as a neutral decision-maker. Likewise, when a viral tweet revealed that the new Apple credit card was systematically giving men higher credit limits than women with identical or better credit, the initial response from the company was to defend "the algorithm" (Webb and Martinuzzi, 2019). When the Stanford Hospital's triage algorithm put administrators who do not deal with the public ahead in line for vaccines compared to residents working in COVID wards, this same pattern of blaming the algorithm was repeated (Guo and Hao, 2021).

In 2019 the US government proposed changes to the Fair Housing Act that would have removed protection against discriminatory outcomes in housing in some cases where algorithms are involved in the decisions. This part of the proposal (which was rejected after public comment) included removing protection for cases where housing decisions that had discriminatory effects were made using a third party algorithm that is "standard in the industry" and being used for its intended purpose. It also included cases where a neutral third party testifies that they have analyzed the model used to make housing decisions, found that its inputs are not proxies for protected characteristics and it "is predictive of risk or another valid objective" (Department of Housing and Urban Development, 2019). Collaborative filtering, as shown here, is standard in its industry, does not use proxies for protected categories, and its objective function, prediction accuracy, is a valid objective, however the algorithm systematically produces discriminatory results. By analogy, housing decisions could likewise be made using algorithms that without being explicitly designed to discriminate, nevertheless do.

We have discussed several types of statistical bias that are inherent in the very logic of collaborative filtering: a class of machine learning algorithm that is in very widespread use. These biases are neither the result of biased datasets,

nor of algorithm builders' personal biases. They are the result of assumptions made in the design of the algorithms themselves. Fixing biased datasets and improving the ethical behaviour of AI workers are also needed steps, but they will not eliminate all sources of bias in machine learning, because there is also bias inherent in algorithms themselves. These are not simply value neutral statistical biases. When marginalized populations are literally on the margins or tails of distributions of user data, statistical biases cause discriminatory outputs.

References

Abdollahpouri H, Mansoury M, Burke R, Mobasher B (2019) The unfairness of popularity bias in recommendation. arXiv:190713286

Adamopoulos P, Tuzhilin A (2014) On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In: Proceedings of the 8th ACM Conference on Recommender systems, ACM, pp 153–160

Ajunwa I, Friedler S, Scheidegger CE, Venkatasubramanian S (2016) Hiring by algorithm: predicting and preventing disparate impact, SSRN

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. Pro Publica May 23, 2016

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp 77–91

Cahn AF, MaGee C, Manis E, Akyol N (2020) Snooping where we sleep: The invasiveness and bias of remote proctoring services. Surveillance Technology Oversight Project November 11, URL https://www.stopspying.org/s/Snooping-Where-We-Sleep-Final.pdf

Cardoso T (2020) Bias behind bars: A globe investigation finds a prison system stacked against black and indigenous inmates. URL https://www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prison-risk-assessments/

Chawla NV, Karakoulas G (2005) Learning from labeled and unlabeled data: An empirical study across techniques and domains. Journal of Artificial Intelligence Research 23:331–366

Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In: IJCAI, vol 17, pp 4691–4697

Department of Housing and Urban Development (2019) FR-6111-P-02 HUD's implementation of the Fair Housing Act's Disparate Impact Standard

Dwoskin E (2018) Men (only) at work: Job ads for construction workers and truck drivers on Facebook discriminated on gender, ACLU alleges. The Washington Post September 18

Edizel B, Bonchi F, Hajian S, Panisson A, Tassa T (2020) Fairecsys: mitigating algorithmic bias in recommender systems. International Journal of Data

Science and Analytics 9(2):197–213, DOI 10.1007/s41060-019-00181-5, URL https://doi.org/10.1007/s41060-019-00181-5

Ekstrand MD, Tian M, Azpiazu IM, Ekstrand JD, Anuyah O, McNeill D, Pera MS (2018a) All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Conference on Fairness, Accountability and Transparency, pp 172–186

Ekstrand MD, Tian M, Kazi MRI, Mehrpouyan H, Kluver D (2018b) Exploring author gender in book rating and recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp 242–250

Evans JA (2008) Electronic publication and the narrowing of science and scholarship. science 321(5887):395–399

Ferryman K, Pitcan M (2018) Fairness in precision medicine. Data & Society 1

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Transactions on Information Systems 14(3):330–347

Fussell S (2017) Why can't this soap dispenser identify dark skin? Gizmodo August 17, URL https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773

Garcia M (2016) Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal 33(4):111–117

Groen D (2018) How we made AI as racist and sexist as humans. The Walrus May. 16, 2018

Guo E, Hao K (2021) This is the Stanford vaccine algorithm that left out frontline doctors. URL https://www.technologyreview.com/2020/12/21/1015303/stanford-vaccine-algorithm/

Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) 22(1):5–53

Hooker S (2021) Moving beyond "algorithmic bias is a data problem". Patterns 2(4)

Kirkpatrick K (2016) Battling algorithmic bias: how do we ensure algorithms treat us fairly? Communications of the ACM 59(10):16–17

LeCun Y (2019) [Twitter]. December 7, (Accessed: October 27, 2020), URL https://twitter.com/ylecun/status/1203211859366576128

LeCun Y (2020) [Twitter]. June 20, (Accessed: Sept 18, 2020), URL https://twitter.com/ylecun/status/1274782757907030016

Meadway J (2020) 'Fuck the Algorithm': How A-level students have shown the future of protest. URL https://novaramedia.com/2020/08/17/fuck-the-algorithm-how-a-level-students-have-shown-future-of-protest/

Mehrotra R, Anderson A, Diaz F, Sharma A, Wallach H, Yilmaz E (2017) Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, pp 626–633

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8(1):141–163, URL https://doi.org/10.1146/annurev-statistics-042720-125902

Neophytou N, Mitra B, Stinson C (forthcoming) Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In: Proceedings of the European Conference on Information Retrieval 2022

Noble SU (2018) Algorithms of oppression: How search engines reinforce racism. NYU Press

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453

Olsen S (2002) Amazon blushes over sex link gaffe. CNET News December 9, 2002

Olteanu A, Castillo C, Diaz F, Kiciman E (2019) Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2:13

Perez S (2015) Apple stops ignoring women's health with iOS 9 HealthKit update, now featuring period tracking. Tech Crunch June 9, URL https://techcrunch.com/2015/06/09/apple-stops-ignoring-womens-health-with-ios-9-healthkit-update-now-featuring-period-tracking/

Raub M (2018) Bots, bias and big data: Artificial intelligence, algorithmic bias and disparate impact liability in hiring practices. Arkansas Law Review 71(2):529–570

Sánchez-Monedero J, Dencik L, Edwards L (2020) What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 458–468

Schein AI, Popescul A, Ungar LH, Pennock DM (2002) Methods and metrics for cold-start recommendations. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp 253–260

Steck H (2011) Item popularity and recommendation accuracy. In: Proceedings of the fifth ACM conference on Recommender systems, ACM, pp 125–132

Stinson CE (2002) Adaptive information filtering with labelled and unlabelled data. Master's Thesis, University of Toronto, Department of Computer Science

Sun W, Nasraoui O, Shafto P (2018) Iterated algorithmic bias in the interactive machine learning process of information filtering. In: KDIR, pp 108–116

Treviranus J (2014) The value of the statistically insignificant. Educause Review 49(1):46–47

Tufekci Z (2018) YouTube, the great radicalizer. The New York Times March 10

Webb A, Martinuzzi E (2019) The Apple card is sexist. Blaming the algorithm is proof. URL https://www.bloomberg.com/opinion/articles/2019-11-11/is-the-apple-and-goldman-sachs-credit-card-sexist

West J (2019) Echo chambers in science?, unpublished manuscript

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, pp 1171–1180