Can an Algorithm be Biased?

Catherine Stinson

Center for Science and Thought, University of Bonn

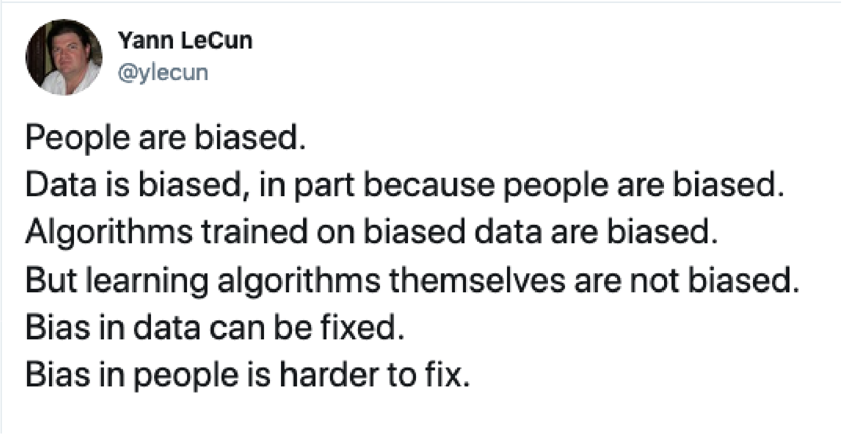
Poppelsdorfer Allee 28, 53113 Bonn, Germany

cstinson@uni-bonn.de

Abstract

Efforts to shine a light on algorithmic bias tend to focus on examples where either the data or the people building the algorithms are biased. This gives the impression that clean data and good intentions could eliminate bias in machine learning. But algorithms themselves are not neutral. This is illustrated with the example of collaborative filtering, which is known to suffer from several statistical biases. Iterative information filtering algorithms in general create a selection bias in the course of learning from user responses to documents that the algorithm recommended. These are not merely statistical biases though; these statistical biases cause "bias of moral import." Marginalized people are literally on the margins of data distributions, as work in disability studies highlights. This source of bias warrants serious attention given the ubiquity of algorithmic decision-making.

Word count: 4794.



Yann LeCun
@ylecun

People are biased.
Data is biased, in part because people are biased.
Algorithms trained on biased data are biased.
But learning algorithms themselves are not biased.
Bias in data can be fixed.
Bias in people is harder to fix.

Figure 1: LeCun (2019)

1 Introduction

A common reaction to claims of algorithmic bias is to dismiss them as category mistakes. “It’s not just a joke anymore: They’re actually claiming math is racist” claims a recent headline (Freddoso, 2017). A more nuanced version is to claim that in cases of algorithmic bias it is always either the people involved, or the data that contribute the bias, not the algorithm itself. The tweet by Yann LeCun, inventor of deep learning, shown in Figure 1, is a case in point.

But there are ways of interpreting ‘Can an algorithm be biased?’ such that it is an interesting question. The first goal here is to clarify the question, sorting out the uninteresting and trivial interpretations from the meatier ones. The second goal is to begin to answer the question by considering arguments and empirical evidence suggesting ways in which algorithms can be biased.

Algorithmic bias is a topic of discussion in machine learning (ML), a branch of artificial intelligence (AI) that has recently expanded its scope to many of the interactions people have with technology. ML is used not only in high-tech tools like self-driving cars and facial recognition, but also in everyday technologies like thermostats and traffic lights. Increasingly ML is being used in less visible contexts like work scheduling, traffic optimization, and search completion. Its use in hospital triage and automated diagnosis makes the question literally one of life or death. If the algorithms employed to make these decisions about human lives are biased, the effects could be massive, but also difficult to discern.

Section 2 breaks down what ‘algorithm’ and ‘bias’ mean, and sorts out the trivial from the interesting interpretations of the question, ‘Can an algorithm be biased?’. Section 3 explores some statistical biases that are well known among ML researchers to affect algorithms. Section 4 argues that these statistical biases can lead to discriminatory outcomes for minorities. Section 5 offers concluding reflections on why it is important not to gloss over bias that derives from algorithms themselves.

2 Clarifying terms

One factor driving disagreements over algorithmic bias is the ambiguity in the meaning of ‘algorithm’. When ‘algorithm’ is understood as just bits of math and logic, it is easy to ridicule the idea that algorithms are biased. Exclusive OR is not elitist, despite the name. The number 55378008 is not to blame for math class sexual harassers’ penchant for showing it upside down on calculators.

Others understand ‘algorithm’ broadly, to describe entire computational systems embedded in social contexts. The Algorithmic Accountability Act introduced to the US Congress in 2019 (H.R.2231, 2019) refers to algorithms in the title, but the text of the bill is about automated decision systems, like the use of facial recognition systems by police forces, or automated hiring systems by corporations. It is just as obvious that these automated decision systems can be biased as it is that “1 + 1 = 2” is not.

Here ‘algorithm’ will be used in an intermediate way, to refer to more than just a set of mathematical or logical operations, but less than an entire computational system embedded in a social context.[1] The dictionary definition of ‘algorithm’ is “A procedure or set of rules used in calculation and problem-solving; (in later use spec.) a precisely defined set of mathematical or logical operations for the performance of a particular task.” (OED, 2020). An algorithm is a set of rules, not an entire computational system, but a set of rules *used for a task*. Alphabetical order, for example, taken as a rule in abstracta may not be the sort of thing that could be biased. Alphabetical order used to organize items into approximately equal sized groups (like the lines at the registration desk at a conference) is not biased, assuming the groups are all treated equally. But alphabetical order used for the distribution of scarce resources would systematically benefit some at the expense of others, and in some contexts even constitute discrimination, given that the names popular in different regions and religions are not evenly distributed across the alphabet. When put into use for a task, the apparent neutrality of algorithms becomes less clear.

2.1 Bias of Moral Import

’Bias’ is still more ambiguous. Friedman and Nissenbaum’s groundbreaking 1996 paper, “Bias in Computer Systems” distinguishes value neutral uses of ‘bias’ from “bias of moral import” or unfair discrimination, which they define as follows: “A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate”

[1] One exception is the phrase ‘algorithmic bias’ which refers broadly to bias in computational systems.

(Friedman and Nissenbaum, 1996, 332).

Bias of moral import they divide into three types, based on an analysis of the source of the bias: preexisting bias, technical bias, and emergent bias. As the name suggests, preexisting bias is at least in part distinguished temporally. These are biases that "exist independently, and usually prior to the creation of the system" (Friedman and Nissenbaum, 1996, 334) either in the society at large, or in individuals involved in the design of the software, and include both conscious and unconscious or implicit biases. Emergent bias also has a partly temporal character. It arises when users interact with the system, and "typically emerges some time after a design is completed" (Friedman and Nissenbaum, 1996, 335). Examples include cases where new knowledge can't be incorporated into the design after the fact, and where users have different abilities, needs, or values than were anticipated by designers, such as giving written instructions to an illiterate population. Technical bias is more of a hodge-podge of examples, including screen size limiting visibility of options, misuse of pseudo-random number generators, imperfect formalization of ambiguous or complex concepts, or algorithms used in contexts for which they are inappropriate.

Friedman and Nissenbaum's taxonomy of bias is not fine-grained enough for contemporary discussions of algorithmic bias. Consider the case of Street Bump, a smartphone app that detects bumps during car rides, then reports the location of the bumps to a central system for allocating road repair resources. As Crawford (2013) points out, the app fails to make allocation of road repair resources fairer, because smartphones are much more likely to be found in wealthier neighborhoods than poorer ones. Is this preexisting bias, because the unequal distribution of technological resources exists independently in society at large? Is it technological bias, because the app depends on technologies only found in certain types of phones? Or is it emergent bias, because there was a mismatch between the actual abilities of users and the abilities anticipated by the designers?

In contemporary discussions, the main lines of division are between biased data, biased people, and biased algorithms (if those exist), as seen in Figure 1. Several cases have been uncovered where using data about past decisions to train systems to make policing, hiring, or credit decisions in the present means that historical discrimination gets trained into the algorithm, perpetuating historical bias in the next generation of decisions (Angwin et al., 2016; Campolo et al., 2017).

There are also documented cases where people have designed algorithms specifically to create discriminatory outcomes. Redlining certain neighbourhoods as high risks for mortgages, based on the racial composition of residents (Gaspaire, 2012), or choosing to target only men to show certain kinds of job ads (Dwoskin, 2018) are two examples. More typical are cases where researchers inadvertently ignore the interests of some groups, such as speech recognition algorithms that fail to work for users with non-standard accents or users recovering from stroke, and automatic soap dispensers that only detect light colored hands. It can be unclear whether to attribute the bias to the algorithm or to its designers.

In other cases it is unclear whether the bias should be attributed to people or to data. That facial recognition algorithms are an order of magnitude less accurate for black female faces than for white male faces is attributed to the lack of black and female faces among the training data used to build facial recognition systems, but this in turn stems from a lack of gender and racial diversity among AI researchers (Buolamwini and Gebru, 2018).

This taxonomy of biased data, people, and (perhaps) algorithms do not exhaust the distinctions one might want to make between types of algorithmic bias. Greater clarity can be found by distinguishing bias along two axes: i) at what point in the design process bias enters the picture, and ii) the source or cause of the bias. Other relevant distinctions not fully

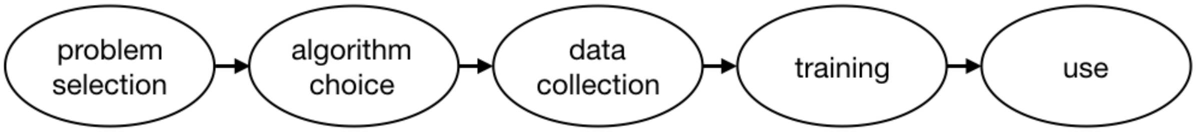


Figure 2: Simplified workflow of a ML system.

addressed here are between biased intentions and biased outcomes, and who (if anyone) is morally blameworthy or legally responsible for a biased system.

Stages in the workflow of a ML system where bias might occur include: problem selection, choice of algorithm, data collection, training, and use. A simplified workflow of these stages is shown in Figure 2. Additional stages like calibration, testing, and re-design may be added.

Sources of bias include people or institutions that either intentionally, or unintentionally cause discriminatory outcomes, as well as what we might call naturally occurring bias. Goods are not evenly distributed among people in the world, not only because of human actions, but also due to accidental or natural occurrences, so data can be biased because of a discriminatory data gathering process, or dumb luck. Explicit bias can drive algorithm choice, but it can also be difficult to predict when an algorithm will be a poor fit. An algorithm might work well for some uses, but produce discriminatory outcomes in other contexts. Particular users or goals may be prioritized in training and testing, while others are overlooked. Mismatches between designers' assumptions and users' needs can happen due to bias on the part of designers, or accidentally.

By conceptually separating the sources of bias from the stage at which it occurs, the question of whether biased people are involved in causing algorithmic bias can be evaluated separately from whether the bias affects the data, the algorithm itself, or some other stage of the workflow. This analysis also helps illuminate why LeCun's suggestion that fixing biased data is sufficient for addressing algorithmic bias is wrong. Bias can get in at later stages too.

In ML ‘bias’ is also used in value neutral ways, to refer to statistical biases. These are explored next.

3 Statistical Biases

In statistics and ML, ‘bias’ has a different set of meanings than Friedman and Nissenbaum’s bias of moral import. Selection bias is when a non-random process is used to select a sample from a population, such that members of the population do not have equal chances of being selected. For example, if people conducting a survey preferentially approach tall people to participate, the resulting sample would be biased, and the survey results might be misleading.

Estimator bias is the extent to which the value of a variable measured in a sample differs from the value of that variable measured in the whole population. For example, if the average height among a sample of people were 210cm, whereas the average height for the population was 170cm, estimator bias would be quite high. Estimator bias can occur as a result of sample bias, or in an unbiased sample.

A related statistic is variance, which measures how far a set of numbers are spread out from their average value. For example, if the average height in a sample were identical to the population average of 170cm, because everyone in the sample was 170cm tall, there would be no estimator bias, and zero variance. But assuming that not everyone in the population is exactly 170cm tall, the population variance would be greater than zero. A mismatch between a sample and a population’s variance is not bias in the statistical sense, but it is another way that a sample can be unrepresentative.

One class of ML algorithms where statistical biases have been well documented is collaborative filtering.

3.1 Bias in Collaborative Filtering

Collaborative filtering algorithms are used in 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...” User profiles are constructed based on both explicit ratings such as likes or stars, and implicit ratings like clicks or viewing time. To come up with recommendations, the user profile is compared to other users’ profiles to find close matches. Items that were rated highly by users with similar profiles are then recommended to the user. User profiles are models of users, and are continuously updated as the user interacts with the system, with the goal of making the profile a more accurate predictor of the user’s preferences.

Some ways in which collaborative filtering algorithms are biased are described below. Olteanu et al. (2019) catalogue a number of additional biases that can occur in the software development cycle.

3.1.1 Cold-Start Problem

The cold-start problem is a well established bias affecting collaborative filtering. The problem is that when a new item becomes available, there are initially no ratings of it by any user. Since recommendations are based on what other users have rated, this means that collaborative filtering cannot recommend new items without a mechanism to counteract this bias.

Furthermore, items that have been in the system longer tend to build up more ratings over time, so are more likely to be recommended than newer items. This dynamic would develop even if initial ratings were evenly distributed, and initial recommendations were randomly sampled. As users interact with the system, they take up some recommendations and not others, so provide ratings unevenly across the space of items. From the perspective of users, the cold-start problem appears as a (small c) conservative bias, where older items are hard to avoid, and new things are harder to find. There is also an implicit assumption that users' ratings will remain constant, rather than their tastes being allowed to change or mature.

3.1.2 Popularity Bias

A related problem is the popularity bias (Herlocker et al., 2004; Steck, 2011), where very popular items get over-recommended. Even a user whose only positive ratings are for medieval Persian editions of ancient medical texts will get recommendations for *Harry Potter*, because no matter what your preferences, there is a good chance that someone with similar tastes liked *Harry Potter*. These are not very useful recommendations for most users. Relatedly, a user might have bought *Fifty Shades of Gray* because they are writing a dissertation about representations of kink in popular culture, then have to wade through pulp romance recommendations popular with its fans, despite having no interest in the genre.

Profile injection attacks take advantage of the popularity bias to manipulate the probability of another item being recommended. 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 (Olsen, 2002).

Many users' preferences will cluster around popular items, but other users might cluster in smaller niche groups (Horror fans, perhaps), and still others will have rare preferences (like the medieval Persian medical text fan), or atypical combinations of preferences (a fan of both Death Metal and musicals, for example). These biases affect users differently, depending on where they are located in the distribution of preferences.

3.1.3 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. A number of papers attribute this problem to an overemphasis on maximizing recommendation accuracy, while overlooking other statistics that affect user satisfaction (Adamopoulos and Tuzhilin, 2014; Ekstrand et al., 2018a,b).

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. For example, a 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*. By choosing a more diverse set of neighbouring user profiles on which to base recommendations, Adamopoulos and Tuzhilin (2014) mitigate both over-specalization and the popularity bias, increasing the diversity of recommendations without sacrificing prediction accuracy.

3.1.4 Homogenization

Homogenization is another issue for which there is scattered evidence. Homogenization is an effect over the dataset as a whole, where the variance of items recommended to all users combined decreases over time.

Since online journals became common, increasing the availability of academic literature, citation practices have narrowed. Fewer articles are being cited, suggesting that people are reading less widely despite greater access (Evans, 2008). A recent study (West, 2019) demonstrates that GoogleScholar's recommendations have also 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. Rather than sampling from the entire dataset equally, this narrowing of recommendations tends to happen when recommender systems only show items that other users liked.

The phenomenon of “filter bubbles” or “echo chambers” is often blamed on the laziness of individuals not bothering to look for media that might challenge their comfortable opinions. However, filter bubbles may result from the homogenization that is characteristic of collaborative filtering algorithms. 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).

3.2 Bias in Information Filtering

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 a user’s preferences or a particular topic. Popular applications include search engines returning pages relevant to a search term, or spam filters quarantining suspicious emails in a folder. *Iterative* information filters continuously update their predictive model based on user feedback to improve performance during operation.

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 feedback labels, which are used to update the model. Then the loop repeats with recommendations based on the updated model.

3.2.1 Selection Bias

The user's interactions change the model, based on what was recommended, which in turn affects what can be recommended at later stages. Iterative information filters (including the subclass of collaborative filtering systems) introduce a selection bias in the course of their operation (Stinson, 2002; Chawla and Karakoulas, 2005). Since labels are much more likely to be provided for items that were recommended, the labeled data form a biased sample. Furthermore, users are more likely to rate items they like than that they do not like, which further biases the sample (Marlin et al., 2012). Yao and Huang (2017) note that "sampled ratings have markedly different properties from the users' true preferences." Sun et al. (2018) report more homogeneous recommendations, and a sacrifice in prediction accuracy compared to unbiased classifiers.

It is utterly uncontroversial among ML researchers that information filtering algorithms manifest a host of statistical biases, so there is one sense in which it is trivial that algorithms can be biased. However the more interesting question is whether algorithms cause bias of moral import. Although social scientists have offered ample evidence that search and recommender algorithms reinforce and amplify inequality (Noble, 2018), the problem could be pinned on biased people or biased data. What remains is to show that statistical bias in algorithms can translate into bias of moral import.

4 From Statistical Bias to Discrimination

Statistical bias has negative effects on the performance of algorithms if uncorrected, which is bad for all users, as well as media producers who want their products to have a fair chance of being seen (Mehrotra et al., 2018). The implications go well beyond occasionally getting unwanted recommendations. As algorithms mediate more and more of our access to information, access to services, and decisions about our lives, their performance becomes a significant equity issue.

Several of the biases described above stem from an over-emphasis on maximizing mean prediction accuracy, and the effect was a tendency to zero in on tastes shared by the majority. This ignores not only the value of “information diversity” (Bozdag, 2013), but also disproportionally disadvantages minority users. This is because minorities are literally on the margins of distributions of human traits (Treviranus, 2014). Designing technologies to work well for the majority clustered around the mean not only disadvantages the other 20% or so of people who occupy the tails of a normal distribution, it wastes an opportunity to make use of the knowledge available on the margins (Treviranus, 2019). Designing technologies (including algorithms) to work well for marginalized users tends to have the side-effect of making them also work better for the average user.

People from minority communities have registered dissatisfaction with search and recommender algorithms. Noble (2018) documents the ways that search algorithms fail to serve the needs of black women. Complaints about culturally inappropriate recommendations, like white hairdressers being recommended for search terms like ‘Black’, ‘relaxer’, and ‘natural’, or Christmas movies being recommended to Jews, are common online. A related complaint 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, like getting recommendations for every coming age story about a gay teen after viewing a single episode of *Rupaul’s Drag Race*.

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 differences in the utility of recommendation systems for users of different demographic groups. Zafar et al. (2017) discuss “disparate mistreatment,” arising when a classifier’s misclassification rates differ across social groups.

There are a number of technological fixes available, like preferentially using items from the tail of a user’s rating distribution as the basis for matching profiles (Steck, 2011). However, these corrections can only be made when we are aware of an algorithm’s biases. False claims about the neutrality of algorithms discourage further research into discovering and fixing bias in algorithms. Perhaps the greatest danger posed by claims that algorithms themselves cannot be biased is that the illusion of neutrality might be exploited in attempts to roll back protections against discrimination.

One example of this is playing out right now in US politics. The Trump administration has proposed changes to the Fair Housing Act that would allow for discriminatory outcomes in housing in some cases where algorithms are involved in the decisions. This includes any cases where a third party algorithm is “standard in the industry” and being used for its intended purpose. It also includes cases where a model “is predictive of risk or another valid objective” (Department of Housing and Urban Development, 2019), and a neutral third party testifies that they have analyzed the model, and found that its inputs are not proxies for protected characteristics. The algorithms described here would pass this proposed Disparate Impact Standard test, despite being biased in ways that can lead to discriminatory outcomes.

5 Conclusions

A non-trivial way of understanding, ‘Can an algorithm be biased?’ is as a question about whether algorithms can cause bias of moral import, independently of any bias that may affect other stages of the ML workflow, like data collection. The answer is a resounding ‘yes’. It is abundantly clear that statistical bias affects many ML algorithms. There is also considerable evidence suggesting that these statistical biases can lead to discrimination. Whether the choice to use a biased algorithm is to be blamed on a biased person or institution, or considered accidental is a separate question. Algorithms themselves can be biased.

Fixing biased datasets and improving the ethical behaviour of AI workers are absolutely necessary steps, but they will not eliminate all sources of bias in ML. The claim that algorithms are neutral is not just false; it is dangerous.

References

"Algorithmic Accountability Act of 2019, H.R.2231, 116th Congress." .

Adamopoulos, Panagiotis, and Alexander Tuzhilin. "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, 2014, 153–160.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *Pro Publica* May 23, 2016 .

Bozdag, Engin. "Bias in algorithmic filtering and personalization." *Ethics and Information Technology* 15, 3 (2013): 209–227.

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*. 2018, 77–91.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. "AI Now 2017 Report." *AI Now Institute at New York University* .

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

Crawford, Kate. "The Hidden Biases of Big Data." *Harvard Business Review* April 1.

Department of Housing and Urban Development. "FR-6111-P-02 HUD's Implementation of the Fair Housing Act's Disparate Impact Standard.", 2019.

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

Ekstrand, Michael D, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. “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*. 2018a, 172–186.

Ekstrand, Michael D, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. “Exploring author gender in book rating and recommendation.” In *Proceedings of the 12th ACM Conference on Recommender Systems*. 2018b, 242–250.

Evans, James A. “Electronic publication and the narrowing of science and scholarship.” *science* 321, 5887 (2008): 395–399.

Freddoso, David. “It’s not just a joke anymore: They’re actually claiming math is racist.” *Washington Examiner* October 24, 2017.

Friedman, Batya, and Helen Nissenbaum. “Bias in Computer Systems.” *ACM Transactions on Information Systems* 14, 3 (1996): 330–347.

Gaspaire, B. “Redlining (1937-).” *https://www.blackpast.org/african-american-history/redlining-1937/* .

Herlocker, Jonathan L, Joseph A Konstan, Loren G Terveen, and John T Riedl. “Evaluating collaborative filtering recommender systems.” *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004): 5–53.

LeCun, Yann. “https://twitter.com/ylecun/status/1203211859366576128.”, 2019.

Marlin, Benjamin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. “Collaborative filtering and the missing at random assumption.” *arXiv preprint:1206.5267* .

Mehrotra, Rishabh, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. “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, 2017, 626–633.

Mehrotra, Rishabh, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. “Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems.” In *Proceedings of the 27th ACM international conference on information and knowledge management*. 2018, 2243–2251.

Noble, Safiya Umoja. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

OED. *algorithm, n.* Oxford University Press, 2020. https://www.oed.com/view/Entry/4959.

Olsen, Stefanie. “Amazon blushes over sex link gaffe.” *CNET News* December 9, 2002 .

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. “Social data: Biases, methodological pitfalls, and ethical boundaries.” *Frontiers in Big Data* 2 (2019): 13.

Steck, Harald. “Item popularity and recommendation accuracy.” In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 2011, 125–132.

Stinson, C. E. “Adaptive Information Filtering with Labelled and Unlabelled Data.” *Master’s Thesis, University of Toronto, Department of Computer Science* .

Sun, Wenlong, Olfa Nasraoui, and Patrick Shafto. "Iterated Algorithmic Bias in the Interactive Machine Learning Process of Information Filtering." In *KDIR*. 2018, 108–116.

Treviranus, Jutta. "The value of the statistically insignificant." *Educause Review* 49, 1 (2014): 46–47.

———. "Inclusive Design: The Bell Curve, the Starburst and the Virtuous Tornado." *Medium* Apr 22.

West, Jevin. "Echo chambers in science?", 2019. Unpublished manuscript.

Yao, Sirui, and Bert Huang. "Beyond parity: Fairness objectives for collaborative filtering." In *Advances in Neural Information Processing Systems*. 2017, 2921–2930.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. "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, 2017, 1171–1180.