Article Citation Format
Dabo, A.M., Sarki, A.M., Fapohunda, S.E. & Longe, E.O. (2022):
A Review of Bias and Discrimination In Digital
Platforms. Journal of Digital Innovations & Contemporary
Research in Science, Engineering & Technology.
Vol. 10, No. 3. Pp 69-82
DOI: dx.doi.org/10.22624/AIMS/DIGITAL/V10N1P5

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received:13th July, 2022
Review Type: Blind
Final Acceptance: 18th September, 2022

A Review of Bias and Discrimination In Digital Platforms

Dabo, Abdulhadi M., Sarki, Abdulrahim Muhammad, Fapohunda Seyi Ebenezer & Longe Edith Osinachi.
Doctoral Programme in Management Information Systems
African Center of Excellence on Technology Enhanced Learning (ACETEL),
National Open University of Nigeria, Abuja, Nigeria
E-mails: abdulrahimmuhammad11@gmail.com; igweseyi@gmail.com; muntaka.dabo@gmail.com; longeedith0@gmail.com
Phones: +2348144864666; +234806 292 4494; +234803639 7682; +2347030146585

ABSTRACT

Digital bias and discrimination have become major concerns around the world, prompting developed and developing countries to set up a legal framework to address the issues; recent articles revealed that Nigeria took a step in this establishment. Users of digital platforms are treated unfairly, unethically, and partially discriminated against based on personal data that is automatically processed by an algorithm. This study used a mixed methods research design. This included conducting a literature search, conducting a systematic literature review by assessing the literature, and analyzing studies on digital discrimination. Digital discrimination often replicates offline discrimination, either by inheriting the biases of previous decision makers or simply reflecting societal prejudices. It may also worsen existing inequalities by giving less weight to historically disadvantaged groups. This review examines this difficult problem, existing solutions, and their limitations from a legal and computer science standpoint. An empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads is featured in the review to highlight algorithmic bias. It concludes with a road map of outstanding issues for mitigating these constraints.

Keywords: Digital Discrimination, algorithmic bias, Machine Learning, AI, Online advertisement

1. INTRODUCTION

Digital discrimination is a type of discrimination in which algorithms use artificial techniques such as machine learning to make decisions that treat users unfairly, unethically, and differently based on their personal data (Such, 2017) such as income, education, gender, age, ethnicity, religion. This is becoming a major issue as reports show that more tasks are being delegated to computers, mobile devices, and autonomous systems. For example, some UK firms base their hiring decisions on automated algorithms. (O'Neil, 2016)

Most nations around the world have anti-discrimination legislation; examples include the International Covenant on Civil and Political Rights, the United States Civil Rights Act, the European Convention for the Protection of Human Rights, the United Kingdom Equality Act 2010, and so on. As in the case of Nigeria, Chairwoman Jessica Rosenworcel has announced the formation of a cross-agency task force to combat digital discrimination and promote equal access to broadband throughout the United States, regardless of zip code, income level, ethnicity, race, religion, or national origin. The Congress also directs the FCC to develop model policies and best practices for states and local governments to use in preventing digital bias in their communities. This critical task will also be overseen by the task force. Furthermore, Congress has requested that the FCC review its consumer complaint process. The task force will also work to improve the agency's approach to gathering feedback from consumers who may be experiencing digital bias in their communities. All of these were approved at the FCC open meeting on March 16, 2022.

However, there is no universally accepted definition of what discrimination is or when it occurs. Indeed, it is a concept that is heavily influenced by culture, social and ethical perceptions, as well as historical and temporal considerations. Most anti-discrimination legislation is simply a non-exhaustive list of criteria or protected attributes (for example, race, gender, or sexual orientation) on which discrimination is prohibited. To put it another way, how much bias is too much? As a result, discrimination is defined as actions or procedures that disadvantage citizens because they belong to specific social groups defined by those characteristics.

2. DIGITAL DISCRIMINATION AS A SUBJECT MATTER

Digital discrimination is defined as "discriminatory treatment or disparate impact acts based on automatic decisions made by algorithms" (Wihbey, 2015). Algorithms, including those that use artificial intelligence techniques such as machine learning, are increasingly being used to make decisions. Sensitive decisions, such as which jobs we apply for, which products we buy, which news we read, and who we date, are increasingly delegated to, or influenced by those systems. Machine learning is a subfield of artificial intelligence that investigates computer algorithms that learn on their own. Unsupervised algorithms (such as those used in data mining) find patterns in each dataset; and supervised algorithms are presented with example inputs and their desired outputs to learn a general rule that maps inputs to outputs.

The target variable is defined as what the supervised algorithm predicts. This target variable can be nominal, in which case its value is referred to as the class category and the machine learning task is referred to as classification, or numeric, in which case the machine learning task is referred to as regression. An algorithm attempting to predict political affiliation from social network data, for example, has a nominal target variable with values representing the various political parties, whereas an algorithm attempting to predict income from purchase data has a numeric target variable. Algorithm-based decisions, including those based on machine learning, are sometimes perceived as faultless, lacking most of the flaws that humans have (e.g., tiredness or personal prejudices); and their decisions may be less scrutinized, that is, decisions made by algorithms are less closely examined than decisions made by humans (Angwin, et al, 2016). Gender inequality persists in most underdeveloped countries, according to reports from the Gender Advisory Board to the Commission on Science and Technology for Development (2006). It was also discovered that countries with more advanced technologies have a lower rate of female internet penetration than their male counterparts. In general, men are found to be more capable of using computers and other ICTs than women.

This included shorter usage periods as well as community access points. According to the report and data collected by the International Telecommunications Union (ITU, 2002), gender patterns in Internet use do not vary equally with Internet penetration, implying that women's rates of Internet use will not automatically rise with national rates of Internet penetration. While the gender gap has closed in countries with high Internet penetration, others, such as Norway, Luxembourg, the United Kingdom, the Netherlands, Germany, and France, have yet to see women's rates of access approach or equal those of men. In fact, for the years for which we have data, women account for less than 40% of Internet users in Germany, the United Kingdom, France, and Norway. On the other hand, the Netherlands' rate of 40% female Internet users is comparable to that of Brazil, Mexico, and Zimbabwe, countries with less than 5% overall internet penetration (Huyer et al, 2005)

2.1 Algorithmic Discrimination

In algorithmic terms, discrimination is simply a deviation from the norm, but it does not always imply unfair treatment of specific social groups. The increased use of algorithms to automate decision making has raised serious concerns that such automated decisions may result in discriminatory outcomes. In settings where advertisements are allocated by algorithms, for example, research compendiums show that instances of historically discriminated-against groups are more likely to be associated with undesirable advertisements and less likely to see desirable advertisements (Sweeney, 2013). (Datta et al. 2015). These studies, however, made no known attempts to learn why these algorithms categories may produce undesirable discriminatory outcomes.

The question was investigated using a set of data obtained from a field study of an advertisement intended to promote STEM job opportunities and training (science, technology, engineering, and math). The emphasis on STEM carriers was motivated by the fact that policymakers in most countries are concerned about a shortage of STEM graduates, particularly women. Because evidence suggests that the shortage is not always the result of hiring practices, providing information about these careers accounts for a sizable portion of this policy challenge. (Williams and Ceci, 2015) demonstrate for an academic context that women are more likely than men to be hired into STEM jobs if they apply. Instead, gender differences in perceptions of STEM careers could explain why women do not apply (Diekman et al. 2010). Thus, disseminating STEM information to women and inspiring them to pursue careers in STEM is an important policy goal (Cheryan et al. 2011, Shapiro and Williams 2012).

The main goal of the advertisement in the field study was to ensure gender neutrality, so the advertiser instructed the ad-serving algorithm to show the advertisement to both men and women. This advertisement was shown in 191 countries around the world. It was noted that the advertisement was seen by more than 20% more men than women. Individuals in their prime career years benefit the most from this distinction. It is commonly assumed that such outcomes occur because those who program the algorithm intend to discriminate or have unconscious biases, or because the algorithm will learn to be biased based on the behavioural data that feeds it (O'Neil 2016). Three explanations were proposed in response to these widely held beliefs.

First Explanation: The discriminatory behaviour was learned by the algorithm from actual consumer behaviour. If women were less likely to click on the ad, an algorithm attempting to maximize click probability might show the ad to men rather than women. However, evidence was presented that showed women were more likely than men to click on the advertisement, ruling out this explanation. A similar explanation could be that there were simply fewer women on the social media platform, for

example, because they spend less time there than men, according to research, making them less likely to see the advertisements. Again, evidence was presented to demonstrate that this was not the case.
Second Explanation: The algorithm learned the behaviour from other data sources on which it was trained, which may reflect a pattern of discrimination against women in various countries. If that were the case, the ad-serving algorithm could simply reflect differences in underlying gender roles in the host country's culture, and the algorithm could have learned over time to present ads in a biased manner. To reflect the level of institutional bias in that country, country-specific data from the World Bank on levels of female education, female labour market participation, or general gender inequality was used. These factors were found to be unrelated to the fact that the STEM advertisement was more likely to be shown to men than to women.

Third Explanation: the algorithm's decision to display the STEM ad less often to women than to men reflected the economics of ad delivery. Multiple advertisers compete for the same set of "eyeballs" in online advertising. Because of the competition, there may be spill over effects from the decisions of other advertisers, even if they are advertising various products. Female eyeballs are more expensive than male eyeballs on average around the world, according to evidence from a separate data collection effort. It was discovered that the price premium an advertiser must pay to show ads to women is especially pronounced for the age group with the strongest negative effect for the display of the STEM ad. Evidence was also provided to support this theory. According to marketing research, women control most household purchases, making them potentially more valuable targets for advertisers. Using data from a separate online retailer, it was then demonstrated that advertisers' higher prices paid for female "clicks" may be profit-maximizing, because women are more likely than men to make a purchase conditional on clicking on the ad.

2.2 Causes of digital discrimination

The current status quo asserts that a level of curiosity about how digital bias occurs may be raised. As illustrated in figure 1, this was explained from the perspective of, but not limited to, modelling, training (online and offline), and the use of machine learning algorithms, which was not intended.

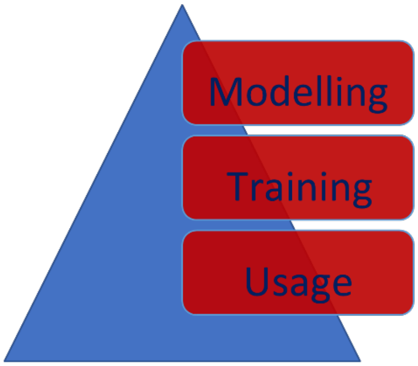


Fig 1: Causes of Digital Discrimination

2.3 Modelling

Machine learning algorithms frequently use modelling to make predictions and recommendations based on specific data. For example, machine learning can define or categorize subjective assessments of a candidate's previous or current success or achievements as a selection or hiring criteria for a candidate. These subjective assessments are certainly biased. Furthermore, if the definition of what qualifies a candidate for employment includes any personal or sensitive information (such as ethnicity) or proxies for this sensitive information (e.g., postcode and income can be a good predictor of race), this can lead to digital discrimination.

2.4 Training (online and offline)

Machine learning algorithms can make effective decisions or predictions based on past decisions from a variety of datasets by training. This dataset could be fed into the machine learning algorithm either offline or online. Machine learning online training enabled them to function as they were instructed to. For example, rewards are given to reinforcement learning algorithms in the same way that punishment is given to bad decisions intuitively (Kaelbling et al, 1996). These systems are always associated with digital bias (inequality). In most cases, online datasets are representative but discriminatory (for example, women, who are more likely to specialize in low-wage sectors, may click more frequently on advertisements for low-wage jobs, which may reinforce the algorithmic rule that suggests these types of job offers to women) (example is some disadvantaged social groups may be excluded from interacting with the data collection system). In the same context, machine learning was shown to acquire prejudices when trained with online texts or through interactions with users (Caliskan et al, 2017).

However, offline machine learning is said to have gone through a series of learning phases in which prediction models were built based on the dataset's given information. It is obvious that discriminatory decisions will be made by the system if the dataset reflects the decision makers' existing prejudices (example, only candidates from a particular group are identified as successful candidates), under-represents a specific social group (e.g., a dataset that does not contain information about a particular social group is likely to lead to inaccurate decisions for that social group), over-represents a specific social group (for example, people who do not conform to stereotypes for a specific profession are usually disproportionately supervised, and their mistakes and flaws are detected at a higher rate), or reflects social inequalities (e.g., particular social groups will have less opportunities to obtain certain qualifications).

2.5 Usage

A machine learning that does not discriminate may be able to do so when exposed to or used in situations for which it was not designed. For example, an algorithm designed to predict the outcome of insect population studies may produce inaccurate results when subjected to different population studies, disadvantage this population over others, and cause discrimination. A natural language processing tool trained on data shared online by one social group may be unable to process online communications produced by another social group, even if both groups speak the same language; for example, research (Eisenstein et al, 2014) suggests that diverse groups use different dialects and word-choices in social media.

3. RESEARCH METHODOLOGY

This review was conducted using a combination of research methods. This method includes a broad search of the literature, a comprehensive review of the literature, data analysis from the literature, and analytical studies of the literature governing digital discrimination. The review emphasizes the detection and avoidance of digital discrimination in the context of machine learning. It also highlights specific cases to illustrate digital discrimination as a topic.

3.1 Research Problem

Many attempts have been made to determine whether the data used to train algorithms bias outcomes are biased or discriminatory. In the medical literature, for example, studies have been conducted with all male research participants. Regardless, the findings have been used to develop diagnostic criteria for whole populations. This has resulted in disparities in detecting heart disease and autism in women. It is problematic to use unrepresentative data to build models or decision-making criteria for entire populations. For instance, researchers (Boulamwini and Gebru, 2018) audited facial recognition software. In response, the companies under investigation implied in public statements that the observed errors in classifying the faces of Black men and women were caused using non-representative data in the development of the technology (Raji and Joy, 2019).

Another effective attempt to avoid discrimination was made by changing how the problem was modelled, so that the protected set of data is not available. For example, research was conducted on the use of machine learning itself, with the goal of identifying recent problem representations that minimize sensitive characteristics inclusion while maintaining affinity for better performance in the trivial prediction task (Edwards & Storkey, 2015; Zemel et al, 2013; Louizo et al, 2015). Although these approaches achieve an acceptable trade-off between prediction accuracy and non-discrimination, it has been demonstrated that the representations learned are not entirely free of sensitive information. Another issue with these approaches is that they cannot handle classification tasks that rely heavily on sensitive characteristics.

3.2 Algorithmic Modifications

Fair machine learning has been reported to expand in recent years from an informed standpoint. Modifications and extensions to existing machine learning and datasets occurred, with the goal of addressing and preventing digital discriminative outcomes. Dwork et al, 2012 conducted one of the earlier works focused on the definition of fair algorithms. Researchers proposed in this paper the development of a classification technique that ensures that similar individuals are treated similarly (regardless of their belonging to protected groups).

This study assumes that there is a task-specific statistic that can be used to compare people. It should be noted that the definition of this metric can incorporate existing decision-making biases about similarity (e.g., not all candidates willing to work longer hours may achieve the same productivity), it can include disparate impacts (or indirect discrimination) to specific groups, and there may not be a clear definition of what constitutes similarity (e.g., how similarities between job applicants can all be defined) (e.g., considering a qualification as a criteria may have a significant negative impact on a particular group of the population less able to obtain it, regardless of their capability to perform well in the job). However, if possible, eliciting a similarity measure may aid in identifying and exposing preconceptions and discrimination that are currently present in decision-making procedures.

The authors (Calders et al, 2013) proposed a method for limiting the impact of unique features in linear regression models. A linear regression model attempts to forecast a numerical result by considering a variety of factors. The authors proposed methods in this paper to ensure that prediction errors for diverse groups are explained by non-protected attributes and that differences in mean prediction in diverse groups, defined in terms of attribute values, are explained by non-protected attributes. In both cases, a definition of protected and nonprotected properties is necessary. Keep in mind that non-protected characteristics may be strongly associated with protected ones and thus serve as a substitute for discrimination.

Raff et al., 2017 proposed a method to reduce discrimination using tree-based algorithms, which are machine learning algorithms that provide some understandability and improve decision transparency. They protected both nominal and numeric protected properties, whereas most existing literature focuses solely on nominal protected attributes.

3.3 Forms of Digital Discrimination

Many researchers established and focused on digital discrimination, which can be highlighted and explained in detail. As illustrated in Figure 2, these include, but not limited to, gender, race, income, location, and lifestyle. Before presenting a generalized case study, this paper examines digital discrimination based on the categories.

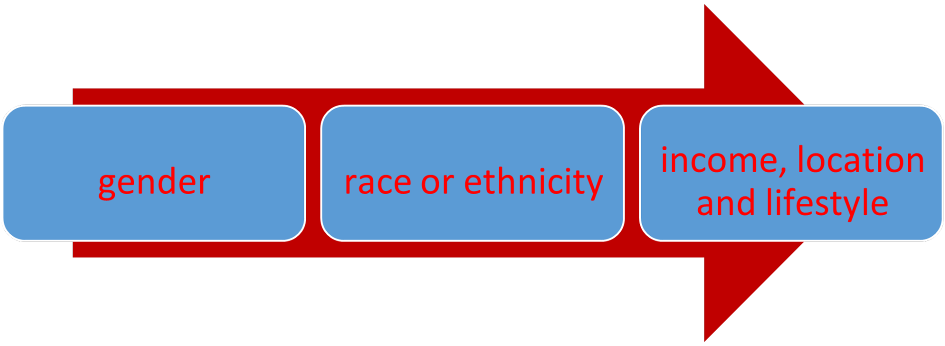


Fig 2: Forms of Digital Discrimination

3.4 Detection and Prevention of Digital Discrimination

The development of metrics and processes to detect discrimination is the first line of defense against discrimination. Discrimination detection metrics can be used in the development and implementation of techniques for avoiding discrimination in the selection of optimization criteria when pre-processing datasets or training algorithms. Within this line of research, Zliobaite (2015) surveyed various metrics proposed to measure indirect discrimination in data and algorithmic decisions.

Other traditional statistical measures that could be used by the researcher to measure discrimination are also discussed in the study. The authors categorize discrimination measures as follows: statistical tests, which indicate the presence of discrimination at the dataset level; absolute measures, which measure the magnitude of discrimination present in a dataset; conditional measures, which capture the extent to which differences between groups are due to protected attributes or other characteristics of individuals; and structural measures, which identify for each individual in the dataset.

Similarly, Tramer et al., 2017 proposed Fair Test, a methodology and toolkit that combines various metrics to detect what they call unwarranted association, which is a strong association between the outputs of a machine learning algorithm and features defining a protected group. Datta et al., 2016, proposed quantitative metrics to assess the degree of influence of inputs on decision-making system outputs. Their research is not meant to detect discrimination in the first place, but the measures they propose have the potential to increase transparency of decisions made by opaque machine learning algorithms, which may provide useful information for discrimination detection.

All these measures assume that the protected ground (i.e., the protected characteristics such as race or gender that cannot be used to make decisions; or the protected groups that cannot be subjected to disparate impact and what constitutes disparate impact) is provided externally, for example, by a law. However, as previously stated, discrimination laws are not exhaustive in terms of all potential instances of discrimination or the grounds for discrimination. As a result of the lack of a clear definition of a protected ground for a given problem, the applicability of these detection measures is limited.

Avoiding digital discrimination is classified according to how the digital discrimination is avoided. It entails modifying the problem model, pre-processing the dataset to be used in the training algorithm, and modifying the algorithm to include non-discrimination as a criterion to maximize alongside prediction accuracy.

3.5 Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career ads

An empirical study of apparent gender-based discrimination in the display of STEM career ads was used to conduct a field test to explain algorithmic discrimination (Lambrecht and Tucker, 2019). The field test was for an advertisement promoting STEM careers. The ad's text was incredibly simple: "Information about STEM careers," accompanied by a picture of the various fields in STEM. Figure 3 depicts a mock-up of a typical ad's ad displays and ad targeting settings. The field test was conducted across three online social media platforms: Facebook, Instagram, and Twitter (Lambrecht and Tucker, 2019).

The same field test was conducted on Google ad woods, Instagram, and Twitter, which are currently the largest social media sites in the world with the highest reach across the United States. Advertisers specify the target audience based on geography, demographics, or interests on such social media platforms, and bid for display advertising impressions to their target audience by specifying a maximum price they are willing to pay per click.

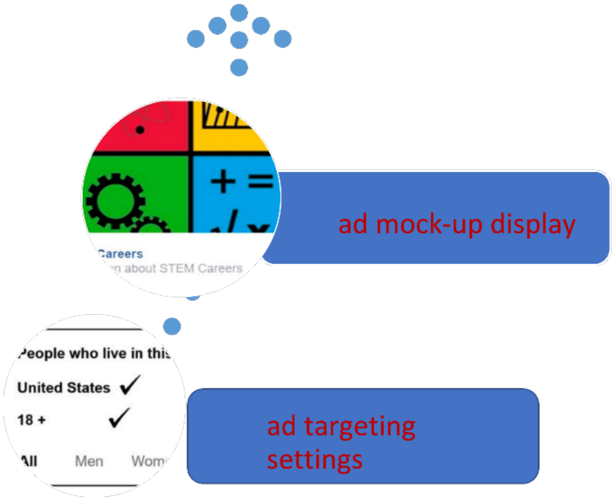
A separate ad campaign with identical ads was created for 191 countries around the world. Later in the article, cross-national variation was used to investigate whether the observed differences in ad allocation can be attributed to different economic and cultural conditions regarding the role of women in different nations. In all cases, the advertisement was aimed at both men and women over the age of eighteen. The only difference between the 191 ad campaigns was the country in which they ran. When a user visits a website, the ad platform typically holds an auction in the background to determine which advertiser will show an ad to that user.

The maximum bid placed by an advertiser in relation to the bids placed by other advertisers usually determined the outcome of the auction. Furthermore, the auction considers an ad's "quality score." The quality score was calculated using a predictive method that assesses the likelihood of a user clicking on any ad (Athey and Nekipelov 2010), adjusting for the relative merit of a bid for the advertising platform. According to Facebook Business (2015), their quality score is referred to as a "relevance score," and "the more positive interactions expected for an ad, the higher the ad's relevance score will be."

Positive indicators vary depending on the objective of the ad, but may include video views, conversions, and so on.)" According to Facebook, the relevance score "can lower the cost of reaching people." Simply put, the higher an ad's relevance score, the lower the cost of delivery." To the advertiser and researcher, the actual calculation of the quality score and the bids of other advertisers that the advertising auction algorithm uses to allocate advertising is a black box (Lambrecht and Tucker, 2019) For all countries, the STEM website initially set a maximum bid per click of $0.20.

The STEM website paused campaigns where the ad had been seen by more than 5,000 people at the end of each day. Due to the delay in pausing the ad campaign, the ad was shown to up to 24,980 users in a country in some cases. If the campaign had not been viewed by 5,000 unique users after a week, the bid was increased to a higher amount that varied by country but could reach $0.60. Bids were received for twenty-nine countries, or 15% of those in the study. These are typically wealthier countries, such as the United Kingdom, the United States, and Switzerland, with higher ad prices.

One concern is that the ad's appearance may elicit different responses from men and women. To investigate this, Amazon's Mechanical Turk was used to determine whether the ad appealed to both men and women. We asked 152 participants from the United States (75 men and 77 women) to assume they saw the ad while browsing the internet and rate their own likelihood of clicking on the ad on a scale of 1 (very unlikely) to 5 (very likely) (highly likely). The average stated likelihood of clicking on the ad was found to be non-significantly different between men (mean 2.053, standard deviation 1.077; p = 0.770) and women (mean 2.105, standard deviation 1.102; p = 0.770). The obtained results were tabulated and discussed in the following sections of this paper.



People who live in this

United States ✓

18 + ✓

All Men Wom

Fig 3: A mock-up of the ad displays and the ad targeting settings for a typical

4. RESULTS

The results obtained from the field test were tabulated below

Table 1: Raw Data Obtained from Facebook Users

Gender	Impressions	Click rates (%)	Total Clicks
Female	1,983,798	0.0084	3507
Male	2,531,798	0.0110	5570

Table 2: Results Obtained from Google Display Networks

Gender	Impressions	Click rates (%)	Cost per click ($)
Female	26,817	1.71	0.20
Male	38,000	0.97	0.19

Table 3: Results Obtained from Instagram Users

Gender	Impressions	Click rates (%)	Cost per click ($)
Female	1,560	0.27	1.74
Male	9.565	0.59	0.9

Table 4: Results Obtained from Twitter Users

Gender	Impressions	Total spends ($)
Female	52,363	31.00
Male	66,243	46.84

4.1 Discussions

From table 1, The ad algorithm revealed that women spent less time on social media than men and had a lower chance of seeing the ad. This is because advertisers have a tough time reaching them. However, according to Facebook internal data, women use fewer social media than men but are more active on platforms than men, and as a result, they are more likely to be exposed to ads. In the United States, women outnumbered men in terms of Facebook usage, with 54% and 46%, respectively. According to data from 2018, men liked eighteen posts and made ten comments on average in the previous month, whereas women liked thirty-six posts and made twenty-nine comments. Men liked nineteen posts and made nineteen comments in an average month, whereas women liked thirty-five posts and made twenty-four comments. In general, 59% of active users on Facebook are women, and 41% are men; however, 44% of Facebook profiles are for women, and 56% are for men. Internal data from Facebook, supported by industry reports and survey evidence. Similarly, according to Vermeren (2015), 76% of women and 66% of men use Facebook, with women having more than twice as many posts on their Facebook wall and 8% more "friends" than men.

Table 2 displayed the Google AdWords campaign statistics The data pattern resembled that obtained from the Facebook test, as shown in table 1. Thirty-six percent of impressions were shown to women, while 51% were shown to men. A further category with unknown gender accounted for 13% of ad impressions. Consistent with previous findings, women were far more likely than men to click on the ad if they saw it. Women were slightly more expensive to advertise to, despite having far higher click-through rates, which should have pushed down pricing.

Table 3 Instagram Test statistics were displayed. Women receive 15% of all impressions. However, Instagram was the only platform studied where men were more likely than women to click on an ad. It is possible that the gender disparity in click rate aggravated the algorithm's ad allocation. The fact that women were far more expensive to advertise to than men support such an interpretation.Finally, Table 4 The results of the Twitter test were illustrated. Because Twitter only reports total spend for each gender group, click rate estimates by gender were not obtained from the Twitter interface. The result obtained for the number of impressions displayed to people, however, echoes that of Table 1 in that, once again, women were less likely to see the ad.

Based on the results from tables 1,2,3, and 4 in section 2.2.1 above, it can be explained that the ad algorithm has learned the preferences of the host countries and knows that it is undesirable to show a specific type of ad, or employment ads in general, to women in a particular country. Discrimination could occur if the algorithm were trained on a training data set that included discrimination or if it had learned discrimination in previous campaigns run by different advertisers. In this case, women's impressions may simply reflect the fact that, in most countries, women's labour market rights and careers lag men.

4.2 Unresolved Issues Associated with Digital Discrimination

Despite the research on digital discrimination, there are still a number of unresolved issues. One special topic of interest for us is how to attest digital discrimination, or test whether new or current algorithms and/or datasets have biases to the extent that they discriminate users. This is a critical step in understanding and avoiding digital discrimination. Even though prior research on detecting digital discrimination can measure and/or limit the degree of bias in an algorithm or dataset, there are still unanswered questions that need to be addressed right away. To put it another way, how much bias is too much? Various measures have been developed by researchers to assess discrimination and unintended correlations between user traits and factors, as shown in Section 2.1.

They do not, however, provide a comprehensive understanding of whether digital discrimination is permissible in terms of law, morality, and/or society. As a result, defining what constitutes digital discrimination and how to translate that into automated procedures that confirm digital discrimination in datasets and algorithms is critical and difficult. This necessitates interdisciplinary collaboration to facilitate a synergistic and transformative social, legal, ethical, and technical approach. This will eventually allow the development of ground-breaking methods for automatically verifying computational non-discrimination norms in datasets and algorithms to certify their discrimination. Two of these unresolved issues are addressed in the sections that follow.

4.3 Socio-economic and Cultural Dimensions of Digital Discrimination

Plane et al, 2017 and Grgic-Hlaca et al, 2018 conducted extensive research on user perceptions of digital bias based on considerations of concrete domains such as online targeted advertisements and criminal risk predictions. The second case, for example, focuses on why people perceive certain features to be fair or unfair when used in algorithms. Although this is a great start and a step in the right direction, more research is needed to lay a solid empirical foundation for understanding the socioeconomic and cultural aspects of digital prejudice. This should include not only survey-based studies, but also the definition of spaces for cross-disciplinary research that draws on the expertise of a diverse research team comprised of experts in machine learning, human-computer interaction, law, ethics, philosophy, social sciences, and other fields. Co-creation spaces, bringing together researchers, technical and non-technical users, should also be defined.

These locations should bring to light the social, economic, and cultural components of digital prejudice by studying datasets, algorithms, and procedures. Techno cultural techniques such as those described in (Cote & Pybus, 2016) appear particularly suited to this application.

4.4 Legal-ethical Frame Works of Digital Discrimination

As previously stated in Section 1.1, many countries already have legislation in place that prohibits discrimination, though it is not always clear what qualifies as a legally protected ground. Furthermore, there is no accepted definition of discrimination that encompasses both what it is and the context in which it occurs. This concept has been influenced by culture, social and ethical norms, and time. Most anti-discrimination laws include a non-exhaustive list of protected characteristics (such as race, gender, and sexual orientation) on which discrimination is prohibited. As a result, establishing a legal and ethical framework to articulate, define, and describe digital discrimination is essential.

This necessitates the development of a new framework for understanding discrimination through the lens of decision-making algorithms. Based on an examination of discrimination law and the socioeconomic and cultural empirical base, a critical analysis of the concept of discrimination and antidiscrimination in a digital context should be conducted. Topics such as intersectionality, positive discrimination, and discrimination vs. freedom of choice should be addressed in the reflection. Through such reflection, an initial set of ethical standards covering any gaps in the law or misunderstandings should be developed. These ethical principles must be methodically challenged and criticized using argumentation methods, situations, and counterexamples. The final aspect of this challenge is to call into question the assumptions that allow non-discrimination norms to be computed, i.e., the conditions that allow an efficient algorithm to determine whether a decision-making algorithm or dataset complies with a non-discrimination norm.

5. CONCLUSIONS

Throughout this review, issues related to digital discrimination were discussed. The legal definition of digital discrimination is highlighted in the review. Preconceptions about well-known cases of digital discrimination were also discussed. The review also included explanations of methods for determining digital discrimination, how it occurs, and, in some cases, how to prevent it. A case study that targeted affected users who are biased by a given algorithm was also described. More intensive research, however, is recommended to address the causes of digital bias, how they aggravate the socioeconomic environment, and to provide effective means of mitigating these problems. Furthermore, as demonstrated by Boulamwini and Gebru (2018), as well as other machine learning researchers, there is a need for a better understanding of what constitutes discrimination.

REFERENCES

1. Altman, A. (2016). Discrimination. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy (Winter 2016 Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/discrimination/
2. Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). Machine Bias: There's Software used Across the Country to Predict Future Criminals and it's biased against Blacks. ProPublica, May 23.
3. Athey S, Nekipelov D (2010) A structural model of sponsored search advertising auctions. Working paper, Stanford University, Stanford, CA
4. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349–4357).
5. Calders, T., Karim, A., Kamiran, F., Ali, W. and Zhang, X. (2013). Control- Ling Attribute Effect in Linear Regression. In Data Mining (icdm), 2013 IEEE 13th International Conference on (Pp.71–80).
6. Caliskan, A., Bryson, J. J. and Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science, 356* (6334), 183–186.
7. Cote, M. and Pybus, J. (2016). Simondon on Datafication. A Techno-Cultural Method. *Digital Culture and Society 2* (2), 75–92.
8. Datta, A., Sen, S. and Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In Security and Privacy (sp), 2016 ieee Symposium on (pp. 598–617).
9. Datta, A., Tschantz, M. C. and Datta, A. (2015). Automated Experiments on ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015 (1), 92–112.
10. Diekman AB, Brown ER, Johnston AM, Clark EK (2010) Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers. Psych. Sci. 21(8):1051–1057
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012). Fairness through Awareness. In Proceedings of the 3rd Innovations in the- Oretical Computer Science Conference (Pp. 214–226).
12. Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of airbnb.com.
13. Edwards, H. and Storkey, A. (2015). Censoring Representations with an Adversary. arXiv preprint arXiv:1511.05897.
14. Eisenstein, J., O'Connor, B., Smith, N. A. and Xing, E. P. (2014). Diffusion of Lexical Change in social media. *PloS one, 9* (11), P113-114.
15. Facebook Business (2015) Showing relevance scores for ads on Facebook. Accessed July 1 2017, https://www.facebook.com business/news/relevance-score
16. Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In WWW.
17. heryan S, Siy JO, Vichayapai M, Drury BJ, Kim S (2011) Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM? Soc. Psychol. Pers. Sci. 2(6):656–664.
18. Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research, 4*, 237– 285.
19. Lambrecht A, Tucker C, Wiertz C (2018) Advertising to early trend propagators: Evidence from Twitter. Marketing Sci. 37(2):177–199.

20. Louizos, C., Swersky, K., Li, Y., Welling, M. and Zemel, R. (2015). The Variational Fair Autoencoder. arXiv preprint arXiv:1511.00830.
21. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Broadway Books.
22. Plane, A. C., Redmiles, E. M., Mazurek, M. L., & Tschantz, M. C. (2017). Exploring user perceptions of discrimination in online targeted advertising. In Unisexx security.
23. Raff, E., Sylvester, J. and Mills, S. (2017). Fair Forests: Regularized Tree Induction to Minimize Model Bias. arXiv Preprint arXiv:1712.08197.
24. Shapiro JR, Williams AM (2012) the role of stereotype threats in undermining girls' and women's performance and interest in TEM fields. Sex Roles 66(3-4):175–183
25. Such, J. M. (2017). Privacy and autonomous systems. In Proceedings of the 26th international joint conference on artificial intelligence (pp. 4761–4767).
26. Sweeney, L. (2013). Discrimination in Online ad Delivery. *Queue, 11* (3), 10.
27. Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Lin, H. (2017). Fairtest: Discovering Unwarranted Associations in Data-Driven Applications. In Security and Privacy (euros&p), 2017 ieee European Symposium on (Pp. 401–416).
28. Valentino-Devries, J., Singer-Vine, J. and Soltani, A. (2012). Websites Vary Prices, Deals Based on users Information. *Wall Street Journal, 10*, 60–68.
29. Vermeren I (2015) Men vs. women: Who is more active on social media? Brandwatch (January 28), https://www.brandwatch .com/blog/men-vs-women-active-social-media/.
30. Wihbey, J. (2015). The possibilities of digital discrimination: Research on e-commerce, algorithms, and big data. Journalist's resource.
31. Williams WM, Ceci SJ (2015) National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. Proc. Natl. Acad. Sci. USA 112(17):5360–5365
32. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. (2013). Learning Fair Representations. In International Conference on Machine Learning (Pp. 325–333).
33. Zliobaite, I. (2015). A Survey on Measuring Indirect Discrimination in Machine Learning.