School of Informatics

Informatics Research Review
Factors that lead to gender discrimination in artificial intelligence

[redacted]

January 2021

Abstract

This review paper provides a new perspective combining actual examples and a step-by-step process of exploration of the causes of AI sexism from three perspectives: academia, industry and commerce. The purpose of the paper is to call attention to gender discrimination in AI and propose solutions at the technical level from improving algorithms and building more balanced data sets, and at the social level from improving the transparency of algorithms by law regulation and education as a fundamental change to the current situation.

Date: Thursday 21st January, 2021

Supervisor: [redacted]

1 Introduction

1.1 background

The original gender inequality occurred when mankind started a more stable agricultural settlements about 5000-8000 years ago in Neolithic Iberia, which caused society dominated by man for a long period [1]. Although the obvious and substantial progress had been made recently, for example, a composite gender equality index covering 129 countries in four fields including health, household, politics and socioeconomic from 1950 to 2003 proved it [2], the gender inequality still exists in some special areas like Saudi Arabia influenced by special cultural contexts that males and females get education in separate institutions [3].
Since the artificial intelligence (AI) was defined and the program which is considered as the first one was presented in Dartmouth Summer Research Project in 1956, which predicted what AI should be like in next 50 years [4], the gender discrimination has become a widespread concern with the advent of artificial intelligence because instead of reducing sexism in society, it has widened the inequality between men and women.
Up to now, most of the articles and studies have only gone so far as to exposing the appearance of sexism in AI and have not thought deeply about the reasons for it. Furthermore, there are also many articles that focus on the study of AI algorithms and data, which contain a lot of complex mathematical formulas, which are highly technical for general readers without computer science or AI background and can only be understood by people within the professional field.

1.2 motivation

Widespread social discrimination existed before the advent of AI, and historical and sociological literature reveal the reasons behind these phenomena, which could also be the real cause of AI bias towards gender, since AI bias may come from inherent human prejudices. In addition, the application areas of AI are so broad that the phenomena and causes of discrimination vary across different fields, the most popular areas will be supported by examples and specific data including language, text and images are chosen to study and discuss. The direct cause of bias affecting AI can be algorithms, data collection, or other parts from the process, and the underlying cause can also be humans' own biases. One of the research motivation is to analyze the reasons and causes for the surface of the gender bias.
It is important that women are at heart of those defining the concept of equity. The other motivation of the research is to call social attention to gender inequality in the field of AI and advance women's career in AI preventing the undermining of advances in gender equality that have been supported by decades of feminist thinking, which is being inspired by the unfair treatment results of women using AI technologies.
This paper will combine social factors of the algorithms and data collection principles of AI to provide an easy to understand and extensive theory to both people with computer science background and those who are not in the professional field. Besides, the paper will focus on ethics, fairness, transparency, collaboration, trust, accountability and morality in AI not technical mathematical knowledge. The motivations mentioned above are concluded with two specific questions that will be discussed and analyzed deeply in the fields of natural language processing (NLP) and facial recognition, which will infer the real reasons for gender discrimination expansion.

1. "What are the direct and underlying causes of gender bias in AI?"

2. How gender bias can be reduced through improvements to AI algorithms and effective measures in the real world

The inherent bias against women represented by AI in academia in general [5][6][7][8][9][10][11], the more direct discrimination against women in industry [12][13][14][15][8][16] and the unfair advertising as well as price discrimination used against women in business [17][18][19][20] provide the evidence and motivation to support this paper. In terms of facial recognition and image processing, data collection is the core part where bias appears often, for example, unbalanced data always lead to unfair judgement, and theoretical basis have been taken into account by recent researches [21][22][23][24]. Also, the new effective terminologies such as debiasing variational autoencoder and adversarial methods combined with CRF were proposed to reduce gender disparities in data balance and algorithm, and the results were better than the current approaches [25][26]. The theoretical guidance of flaws in the NLP has been introduced in [14], and researchers' analysis of NLP algorithms to suggest solutions that can effectively reduce gender discrimination, one of which is loss function combined with CDA [27].

2 Paper organization

This section provides the reader with a brief overview of the structure and gives a clear guide to the main topic or subtopic in each section. The whole paper consisting of four sections which are introduction, fields that gender discrimination exists, reasons and causes, improvements, and conclusion.

- Section 1 introduction contains background and motivation. Background describes historical roots of gender discrimination, the social reality of gender discrimination in recent years and the concept of artificial intelligence as well as the development of AI. Also, the limitations of the previous articles are revealed briefly in this part. Motivation is the essential part which explains two questions related to the main topic and appropriate paper selection.

- Section 2 fields that gender discrimination exist has three subsections which are academia, job marketing and crime detection, which attempt to describe examples of AI sexism in various areas of life to illustrate how widespread this phenomenon is.

- Section 3 reasons and causes consists of two subsections including data collection bias and inherent human prejudices. There are many potential parts that can lead to AI sexism and these factors will be pointed out in this section.

- Section 4 Improvement proposes specific measures containing building a fairer data set, improving algorithmic transparency and education development to address each of the possible causes of AI gender discrimination analysed in the previous section.

- Section 5 is the conclusion.

The connections between each section are organized tightly to capture the main topic, and progresses from the current phenomenon of gender discrimination to the imbalance in algorithms and data collection behind it and to deep bias from human being, exploring the most fundamental causes. AI is a mature technology field and AI discrimination is a widely discussed issue, with many articles providing theoretical support and data evidence. The purpose of the paper is to provide a fresh perspective by surveying previous researches in a simpler and more accessible way, and to remedy and improve on the gaps of previous researches.

3 Fields that gender discrimination exist

The truth is that AI is sweeping into every aspect of our lives such as shopping, job hunting, prisoner trials, etc. In terms of gender, the fact that AI is currently a male-dominated field has made many AI genetically sexist. This section will describe examples of gender discrimination related to AI in different fields.

3.1 Academia

Gender discrimination and prejudice are deeply entrenched in academia, even in the West, where awareness of gender equality is higher and mechanisms are relatively more robust. "Experiencing a hostile departmental climate, feeling isolated and invisible, and encountering little or no transparency in departmental decision making facilitates conditions that increase the likelihood of a woman leaking from the pipeline before, during and after tenure decisions are made." shows there are inherent systemic disadvantages of women in the academic field [5]. This is reflected, for example, on the fact that female teachers are less likely to be promoted in the teaching system than male teachers and receive more negative feedback on student evaluation from mixed teaching teams [6]. Also, Women PhDs in chemistry face more barriers to their studies, such as lack of mentorship, feeling isolated, and the need to make extra sacrifices on top of their scientific work to meet the traditional gender division [7].

Widespread sexism in academia has led to women in AI field not being except and even more serious. Compared to organic science, women are still very underrepresented in the research of Technology, Engineering, and Mathematics Education. At the time of writing, only 18% of the world's top AI academic conferences are attended by female authors; only 20% of global professors in AI are women [8]. And the UK is one of the worst areas with less than 10% female engineering professionals in Europe compared to the maximum rate of 30% [9]. According to the government figures, the bachelor's degrees in engineering earned by male are almost four times than those earned by female conferred by post-secondary institutions in 2017-18 in the US [10]. Because of the inherent patriarchy in academia, coupled with stereotypes that men are perceived to be better suited for STEM professions, which is actually wrong [11], resulting in women receiving unfair treatment of the academic field of AI.

3.2 Industry

Compared to academia, the gender bias in industry is reflected more frequently and directly in the imbalance between the product itself and the gender ratio of those working in it. The most obvious recent example is that the use of the voice assistants has risen dramatically during the lockdown in the UK caused by COVID-19 pandemic. Most of them use a female voice as the default option by the result that people treat computer differently and like female voice more than male voice [12] such as Alexa and Siri which default to "female" for most languages setting. However, they have been figured out that they have learned gender bias by using word-embedding technology and female voice is the representative of the submission and compliance, which is prejudice against women and also the reason for the widespread use of female voices [13][14].

Faced with the massive amount of job seekers' resumes, many technology companies have started to use machine screening to replace HR's manual screening. Among them, Amazon (http://Amazon.com), the largest US e-commerce brand, used AI algorithms to initially screen candidates' CVs when making employee hires since 2014 score resumes to reduce HR's work. However, in a Reuters article in 2018 [15], it was revealed that the CV screening techniques used by Amazon during 2014-2015 were more male-biased. The reason for this was actually quite simple, as the engineers trained the algorithm with the CVs of employees who had already been hired by Amazon before. However, this makes Amazon's CV screening algorithm more biased towards male employees. Besides, as similar as the low proportion of academia, the same urgent situation is happening to two tech giants, Facebook and Google, as only 15% and 10% of the AI research teams respectively, are women [8]. Even though women achieve nothing less than, or even better than, men in engineering disciplines, for example, 79.8% of female engineering students achieve a second or first class degree, compared to 74.6% of male students, the number of women working in engineering sector was far from what we expected above, in 2018 only 12.37% of engineers were women in the UK and 21.80% of women were working in the engineering sector. [16].

3.3 Target advertising and price discrimination

For advertisers, AI analyses users' browsing behaviour, consumption habits and other data to push relevant content to the audiences who are most likely to interact with it, which helps advertisers to reach precise audiences more quickly. But for customers, some of the most common forms of AI bias follow such as sexism in targeted advertising, as well as price discrimination behind big-data analysis to price products against existing customers. The gender discrimination is mainly reflected in loss of objectivity and neutrality in the production and distribution by programs, which "skews" the public's objective and comprehensive perception of information. This skewed adv delivery was proved to be used in Facebook which allows advertisers to attract certain users and exclude others by gender. For example, automatic advertisement system of Facebook delivers more job ads for nurses and secretaries to women while lumber and AI related to men [17]. Based on the wealth of data provided by users and inferred from their online activity, this powerful targeting is the key to the popularity of Facebook advertising, which accounts for the majority of Facebook's revenue. Besides, Google's ads was also demonstrated to be skewed such as showing gender discrimination and a lack of transparency through a tracking tool called AdFisher. The data showed that when Google determined that a visitor was a male job seeker, it was far more likely to push an ad for a high-paying executive position to him than to a female job seeker of the same profile [18], which will increase income inequality between men and women.

Price discrimination is originally a basic concept in the economic field and occurs generally when two different people receive different prices for the same product. Widespread e-commerce platforms have taken advantage of AI differentiated pricing system with big data management and the information opacity to engage in hidden price discrimination in order to make profits, which amplifies the information asymmetry that already exists and leads to the violation of consumers' right to fair trade, information, and even privacy. A study from Northeastern University measured price discrimination among 16 e-commerce platforms and found that 9 out of 16 altered prices for different customers especially travel sites [19]. Similarly, dynamic pricing may put women at a greater disadvantage than men when shopping online. According to a study by Department of Consumer Affairs (DCA) in New York City, the women's product cost more than similar products of men such as personal care products and home health care products, which are 13% and 8% higher separately [20]. The fact showed above are likely to be exploited by e-commerce sites to raise prices for female customers.

4 reasons and causes

AI decision-making relies on the learning of human decision-making preferences and outcomes, and machine bias essentially projects biases rooted in social traditions. These are learned from user interactions and re-presented nakedly to a wider audience by AI products, thus creating a chain of bias cycles. Machines never create biases independently, and so-called discrimination and bias are learned from several important aspects of machine learning including datasets and human bias itself, which will be discussed in detail in this section.

4.1 data collection bias

One of the main drivers of bias in AI is training data. Most machine learning tasks are based on large, annotated datasets. If the data set itself is not representative, it does not objectively reflect reality which leads to inevitably unfair algorithmic decisions. A common example is rationing bias, where datasets tend to favour more 'mainstream', accessible groups due to the ease of data collection, and are thus unevenly distributed at gender levels. This fact was confirmed in the image generation study where results that there was a bias towards females in the younger age groups can be explained by the training data containing more samples of young women, which inferred the output distribution depends on the input training data [21]. In simpler terms, the prediction results of machine learning are always biased towards the category with the highest number in the training set. The most well-known training dataset for image recognition, Google ImageNet with over 1.2 million images was found that images assigned to bridegroom from the US and Australia had higher accuracy and confidence than those from developing countries. A third of Google ImageNet were from the US and over 60% from the top 6 European and American countries. China and India, the two countries with the world's largest populations, but only a mere 3% of the data in the dataset [22]. The accuracy (Equation 1) is meaningless if there is a severe skew in the class distributions [23], where FP means false positives – for example, non-spam emails mistakenly classified as spam, FN means false negatives – for example, spam emails mistakenly classified as non-spam and TP/TN means true positives/negatives – correctly classified spam/non-spam.

$$Accuracy = (1 - error) = \frac{correct}{total} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Considering a case where AI determines whether a person is male or female, and the data set used to train the model consisted of 98% males and only 2% females. In the case of predicting whether a less probable event will occur or not, as well as the resume screening system designed by Amazon mentioned in the previous section, the model prefers to predict all the testing class as the same label as the majority class in the training set. The trained AI would be 98% accurate even if it could identify all men, regardless of the 2% of women, which would be very detrimental to the minority in the data.

The model determines element with the highest probability of occurrence, which based on the association with the labels while extracts the labels of the elements in an image directly by identifying the features of the image. That is a very common idea in large information-intensive image recognition. Even in an overall gender-balanced dataset, the choice of data is also important. The degree to which labels are tied to gender can affect predictions such as standing in the kitchen, doing housework, looking after small children are assumed to be female, while meeting, working, playing sports are assumed to be male. Two representative image training datasets supported by Microsoft and Facebook, ImSitu and MSCOCO, each containing over 100 thousand images were found that more than 45% of verbs and 37% of nouns show a gender bias. In the ImSitu where labels were prominently tied to gender, cooking" was associated with women 66% of the time, compared to 33% for men; however, the AI trained for this dataset predicted that cooking was associated with women 84% of the time, compared to only 16% for men [22]. The problem that it showed a greater bias than the dataset itself is that if the recognition is based on existing associations, then the machine may exaggerate the existing associations in the training, thus giving a result that is more likely to be close to the "correct answer" in a less certain circumstance. This is known as 'overfitting' that occurs during the process when model is trained. The models can achieve very high accuracy in the training set, but low accuracy in testing set which inevitably leads to various biases in reality [24]. In fact, the data set behind almost every machine learning algorithm is biased.

4.2 inherent human prejudices

Algorithms are a mirror that reflects many of the biases inherent in human society. Natural Language Processing (NLP) is used as the core part of voice assistant systems such as Amazon's Alexa and Apple's Siri, showed a gender bias in word-embedding, where these systems often outputs that "a father is to a doctor as a mother is to a nurse". Clearly, such associations undoubtedly reflect outdated notions of a modern society [14]. AI algorithms essentially learn from historical human behaviour and decision-making, and this notion of association in AI is undoubtedly an inheritance of the gender bias of human society. Not only does AI's gender bias reflect the gender stereotypes and biases that exist on society, but AI further amplifies these biases in the process of designing and marketing decisions. If not enough women are involved in the sample, then there are gaps between the AI's knowledge, which is why gender bias can occur and humans' own biases will be involved into the AI system.

One of the positions where women are missing is algorithm engineer. They are involved in the whole system including setting the objectives of the machine learning, the choice of model and data labels, pre-processing of the data, etc. Although, inappropriate setting may bring bias from the beginning like trying to identify criminals by their faces, but a more common substitution of personal bias occurs when data features are selected. So the engineer sets labels on the data set to determine what algorithm will learn within that data set and what kind of model it will generate. Amazon's CV screening mechanism for candidates, mentioned in a previous section, shows that if a company employs significantly more men than women, AI is likely to correlate the attractiveness of candidates with certain factors for male candidates, or simply eliminate those applications that have factors associated with female candidates [15]. As algorithmic engineers are heavily dominated by men not only in academia but industry, gender was considered as an important criterion [8][9][16]. The algorithm identifies this particular attribute and builds models around it when learning past hiring decisions, which undoubtedly affects how the algorithm reacts to the data.

For some unstructured data sets (e.g. large amounts of descriptive text, images, videos), the algorithm is unable to analyse them directly. This is where a data annotator is needed to annotate the data and distil the structured dimensions that can be used to train the algorithm. As a simple example, sometimes Google Photos will ask you to help determine which image contains a vehicle, and you are then involved in tagging this image. The data annotator as the processor of the data are often asked to make subjective value judgements, which in turn can be a source of bias. ImageNet as the world's largest database for image recognition is a case in point where many of the images on the site are manually annotated with a variety of segmentation tags [22]. The way images are labelled is a product of the human worldview, and any classification system will reflect the values of the classifier. Prejudices about different cultures and ethnicities exist in different cultural contexts.

5 Improvements

The unknowable and untraceable algorithmic bias makes the improvements a difficult one. Under the existing response system, technological breakthroughs, and policy regulation have tried to address this problem from different aspects.

5.1 Algorithm and dataset improvements

If the dataset used to train a machine learning algorithm is not representative of objective reality, the results of the algorithm from application are often discriminatory and biased against specific groups such as women. Therefore, the most straightforward solution to algorithmic bias is to address it at a technical level, for example by adapting an unbalanced dataset or improving an existing algorithm.
Compared to traditional approaches CDA and REG, the new term of the loss function (Equation 2) combined with CDA was shown to have better performance in reducing gender bias for NLP applications.

$$L = \frac{1}{T}\sum_{t=1}^{T} L^{CE}(t) + \lambda L^{B}(t) \tag{2}$$

According to the data obtained from the experiments where B^N and B^N_c (smaller B^N and B^N_c indicate a more significant reduction in bias) were used to described bias metrics, the bias is significantly mitigated more when the value of λ in the loss function is increased from 0.01 to 2, and the optimum is reached when $\lambda = 1$. The B^N and B^N_c were 0.205 and 0.145 separately, which were smaller than B^N and B^N_c measured when using CDA or REG only. The result revealed that the effect of combining CDA and the loss function where value of λ is 0.5 is more better [27].

In the field of image processing and recognition, methods for autonomous testing of datasets are proposed. For example, debiasing variational autoencoder (DB-VEA) is a type of unsupervised learning can automatically remove data bias from an AI system by resampling it. The model learns not only facial features such as skin, colour and hair but also gender and age, which brought a significant improvement in classification accuracy and reduced bias against gender and ethnicity [26]. Also, the definition of dataset leakage (Equation 3) and model leakage (Equation 4) were came up with to show gender bias amplification (Equation 5) in the same accuracy.

$$\lambda_D(a) = \frac{1}{|D|}\sum_{(Y_i,g_i)\epsilon D} \mathbb{1}[f(r(Y_i,a)) == g_i] \tag{3}$$

$$\lambda_M(a) = \frac{1}{|D|}\sum_{(\hat{Y}_i,g_i)\epsilon D} \mathbb{1}[f(\hat{Y}_i) == g_i] \tag{4}$$

$$\Delta = \lambda_M(a) - \lambda_D(a) \tag{5}$$

where $\mathbb{1}$ is the indicator function and D is the annotated dataset.

The combination of Adversarial Methods (adv) and conditional random fields (CRF) was shown to reduce over 60% of bias amplification on COCO dataset compared to other two CRF-based models including original CRF and CRF combined Reducing Bias Amplification (RBA) [25].

Correcting data proportions is another way ensuring fairness in decision making with a fairer data source. Microsoft cooperated with experts to correct and expand the data set which trained the Face API through Azure Cognitive Services, which provides pre-trained algorithms to detect, recognise and analyse attributes in face images. The new data reduces the recognition error rate of women by 9 times by adjusting the ratio of skin color, gender and age. This shows that building more impartial data sets is undoubtedly one of the basic solutions to the bias occurrence in algorithm part, which is the direction that many companies and academics are working towards.

5.2 Regulation

Promoting algorithmic transparency through policy regulation and ethical guidelines is an effective way to avoid gender discrimination. The EU General Data Protection Regulation (GDPR), which was introduced in 2018 and prohibits automated individual decision-making from being based on special categories of personal data in Article 9 including ”racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation” [28]. The GDPR clarified the rules and responsibilities for online service providers in the collection and use of personal data of European users. The UK Government's updated Data Ethics Framework on 30 August 2018, which requires algorithms to have a certain level of transparency, accountability and fairness in the work with data and AI during the entire process. Transparency is the most important attribute from the overarching principles, which means data and actions must be published openly in a complete format [29]. Google is also responding to the need for greater transparency by proposing Model Cards, which are evaluation reports as similar as algorithm manuals that explain the algorithm used, its strengths and limitations, and even the results of its calculations in different datasets [30].

5.3 Education

Currently, there is one mainstream solution to reduce gender bias, which is to consciously increase the diversity of the AI workforce by adding more women to the AI field. However, this approach has had little effect and does not address the underlying causes at least for the time being. For example, even if the proportion of women in the field of AI increases, there will still be problems such as women seldom entering senior management positions in AI-related companies, and inequality between men and women such as gender pay. It is difficult for women to compete with men because so few women have a true grasp of the core skills and essentials of AI, even if gender discrimination in workplace is removed. The issue of sexism present in companies is one of the reference elements that affect women's choice of computer science profession. In other words, the less respectful technology companies treat women and the fewer opportunities they give to women, the worse women's impression of this industry and the fewer women will choose computer and related careers. The imbalance between men and women in technology companies is mainly due to the imbalance between men and women in science and technology education. According to the data shown in 3.1 section, women are very underrepresented in AI academia, especially in the UK [8][9]. What tech companies need to do is not necessarily to ensure an absolute gender balance, but to keep it in a healthy range. There should also be more diversity of the educational aspect of AI so that more social groups can join in AI learning.

6 Conclusion and Summary

This review paper draws the following conclusions from examples of actual occurrences of AI sexism in academia, industry and business:

- The AI field has inherited inherent discrimination against women in academia leading to an imbalance in the high proportion of male and female professors.
- The perceived submissive voice of women has been widely used in commercial products resulting in a very low percentage of female employees in companies, especially engineers.
- Women in business are less likely to receive well-paid advertising and tend to pay more for the same goods. Earning less and spending more results in a significant imbalance in the distribution of wealth.

In exploring the technical reasons for gender discrimination in AI, it was found that imbalanced datasets which were predominantly male than female, and the biased data labelling were often the causes of gender discrimination. By examining the existing researches, the combination of loss function and CDA methods was found to be effective in reducing gender bias in the field of NLP and the combination of Adversarial Methods (adv) and conditional random fields (CRF) was found to reduce gender bias amplification. These reasons come from the social unfriendliness and disapproval of women specifically in the form of the lack of women in engineer and label annotator positions, which can be solved by improving algorithmic transparency by law regulation and changing education.

References

[1] Marta Cintas-Peña and Leonardo García Sanjuán. Gender Inequalities in Neolithic Iberia: A Multi-Proxy Approach. *European Journal of Archaeology*, 22(4):499–522, nov 2019.

[2] Selin Dilli, Sarah G. Carmichael, and Auke Rijpma. Introducing the Historical Gender Equality Index. *Feminist Economics*, 25(1):31–57, jan 2019.

[3] Ibrahim Mutambik, John Lee, and Abdullah Almuqrin. Role of gender and social context in readiness for e-learning in Saudi high schools. *Distance Education*, pages 1–25, sep 2020.

[4] Minsky M. L. Rochester N. Shannon C.E. McCarthy, J. A proposal for dartmouth summer research project on artificial intelligence. Dartmouth College in Hanover, New Hampshire, 8 1955.

[5] Courtney E. Gasser and Katharine S. Shaffer. Career Development of Women in Academia: Traversing the Leaky Pipeline. *The Professional Counselor*, 4(4):332–352, oct 2014.

[6] Natascha Wagner, Matthias Rieger, and Katherine Voorvelt. Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, 54:79–94, oct 2016.

[7] Jessica UK Lober Newsome for the Resource Centre for Women in SET and the Royal Society of Chemistry. The chemistry PhD: the impact on women's retention. Technical report.

[8] Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark Openai, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz, and Zoe Bauer. Steering Committee. Technical report, 2018.

[9] Jemima Kiss. Vince cable says uk economy hampered by lack of female engineers. *The Guardian*, 2013.

[10] The Condition of Education - Postsecondary Education - Programs, Courses, and Completions - Undergraduate Degree Fields - Indicator May (2020).

[11] Gijsbert Stoet and David C. Geary. The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, 29(4):581–593, apr 2018.

[12] Nass C. I. Reeves, B. *The media equation: How people treat computers, television, and new media like real people and places.* Center for the Study of Language and Information; Cambridge University Press, 1996.

[13] Rebecca Kraut Mark West and Han Ei Chew. *I'd blush if I could closing gender divides in digital skills through education.* EQUALS and UNESCO, 2019.

[14] Tolga Bolukbasi, Kai Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4356–4364. Neural information processing systems foundation, jul 2016.

[15] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[16] Stephanie Neave, Gemma Wood, Tom May Research, Martina Tortis, Maiju Kähärä, Research Assistant, Robin Mellors-Bourne, Maya Desai, Caroline Roberts, Charles Parker, and Susan Wilkinson. Engineering UK 2018: The state of engineering. Technical report.

[17] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), apr 2019.

[18] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, apr 2015.

[19] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring Price Discrimination and Steering on E-commerce Web Sites.

[20] Julie Menin. A Study of Gender Pricing in New York City. Technical report, 2015.

[21] Joni Salminen, Bernard J Jansen, Soon-Gyo Jung, and Shammur Chowdhury. Analyzing Demographic Bias in Artificially Generated Facial Pictures. 2020.

[22] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. Technical report.

[23] Paula Branco, Luis Torgo, and Rita Ribeiro. A survey of predictive modelling under imbalanced distributions, 2015.

[24] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1521–1528. IEEE Computer Society, 2011.

[25] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Technical report, 2019.

[26] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Daniela Rus, and Sangeeta N Bhatia. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, volume 19, New York, NY, USA, 2019. ACM.

[27] Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, pages 223–228, may 2019.

[28] Regulations regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, L 119:38–39, 2016-04-27.

[29] UK Government. Data Ethics Framework. Technical report, 2020-09-16.

[30] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 220–229, oct 2018.