AI & SOCIETY (2024) 39:2045–2057
https://doi.org/10.1007/s00146-023-01675-4

OPEN FORUM

Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities

Sinead O'Connor[1] · **Helen Liu**[1]

Received: 3 February 2023 / Accepted: 11 April 2023 / Published online: 9 May 2023

Abstract
Across the world, artificial intelligence (AI) technologies are being more widely employed in public sector decision-making and processes as a supposedly neutral and an efficient method for optimizing delivery of services. However, the deployment of these technologies has also prompted investigation into the potentially unanticipated consequences of their introduction, to both positive and negative ends. This paper chooses to focus specifically on the relationship between gender bias and AI, exploring claims of the neutrality of such technologies and how its understanding of bias could influence policy and outcomes. Building on a rich seam of literature from both technological and sociological fields, this article constructs an original framework through which to analyse both the perpetuation and mitigation of gender biases, choosing to categorize AI technologies based on whether their input is text or images. Through the close analysis and pairing of four case studies, the paper thus unites two often disparate approaches to the investigation of bias in technology, revealing the large and varied potential for AI to echo and even amplify existing human bias, while acknowledging the important role AI itself can play in reducing or reversing these effects. The conclusion calls for further collaboration between scholars from the worlds of technology, gender studies and public policy in fully exploring algorithmic accountability as well as in accurately and transparently exploring the potential consequences of the introduction of AI technologies.

Keywords Gender · AI · Public policy · New technologies

1 Introduction

Governments often profess their desire to eliminate bias in their products and services, whether that be through implementing affirmative action policies or providing employee training**.** In recent years as discussion around issues and concepts of gender has increased in the public arena, the issue of gender bias in public (and private) institutions has come to the fore. At the same time, governments are incorporating new technologies such as AI into public policy, leading to faster delivery of services, reduction of costs, increased accuracy and new capabilities (Filgueiras 2022:1474). Despite this, when discussing the implementation of such technologies, public policy researchers rarely consider the possible gendered "threats and benefits" of such adoption (Feeney and Fusi 2021:116).

Orlikowski posits that technologies are "products of their time and organizational context" which "will reflect the knowledge, materials, interests, and conditions at a given locus in history" (1992:421). As technology is "both structurally and socially constructed", it both mirrors the implicit biases of its creators, while also gaining new meanings and functions—and potentially biases—through repeated and widespread use (1992:403). When governments employ these technologies, from search engines to recruitment software, they may be unwittingly amplifying such biases, which in turn may influence outcomes from policy to hiring decisions.

Many examples already exist of the automation of decision-making processes in the public sector, such as crime prediction and policing decisions across the US, Netherlands and the UK (Busuioc M 2021:826). As such, several recent studies have called for the public administration field to proactively focus on a research agenda for the introduction of these new technologies (Agarwal 2018) with others also advocating for the inclusion of a feminist perspective (Feeney and Fusi 2021; Savoldi et al. 2021). Now is a critical

✉ Helen Liu
helenliu4@ntu.edu.tw

[1] Department of Political Science, National Taiwan University, Taipei, Taiwan ROC

juncture for scholars from all disciplines to further explore and seek to more fully understand the potential discriminatory effects of AI across different aspects of our lives.

The purpose of this study is to understand the impacts of AI technologies on gender biases. To systematically examine this issue, we develop an analytical framework with two dimensions, namely the direction of gender bias and AI data sources. The direction of gender bias dimension includes gender bias perpetuation and mitigation, while in the dimension of AI functions are images and text.

Due to restraints of time and space, this study will not focus on intersectional biases, e.g. where multiple biases are present simultaneously (for example, gender and disability bias, gender and racial bias etc.), choosing to focus on gender bias alone.

This paper aims to look at just one type of algorithmic bias—gender bias. The study will explore in-depth case studies on four examples of both gender bias *perpetuation* and gender bias *mitigation* found in AI technologies. From these case studies, this paper will describe the existence of gender bias in algorithms, as well as how algorithms themselves could contribute to mitigating gender bias. Finally, this paper will suggest lessons which can be learned from these case studies going forward.

Research questions

1. What current gender biases exist or are enhanced by AI technologies?
2. What gender biases can be reduced by AI technologies?
3. What are the examples that gender biases are created or enhanced by AI technologies?

2 AI and gender biases

There is already a large body of literature which looks at the relationship between humans and technology. Orlikowksi's seminal 1992 work introduces the concept of the 'duality of technology' to express how technology is "physically constructed by actors working in a given social context, and technology is socially constructed by actors through the different meanings they attach to it and the various features they emphasize and use" (406). She posits that the repeated and reflexive mutual interaction between human agents and technology constitutes technology's role in society.

Fountain's 2004 work on information technology and institutional change similarly emphasizes the mutually reinforcing effects of technology and human agency, but places this in an organizational and institutional context. They suggest that the reciprocal influences of organizational/institutional arrangements and information technologies on one another are such that "the effects of the Internet on government will be played out in unexpected ways, profoundly influenced by organizational, political and institutional logics" (12). As part of their aim to plug the gap between the "importance of the Internet and its effects on government and society and the attention of social scientists to this empirical phenomenon", they develop an empirical framework of "technology enactment" which "extend[s] institutional perspectives to account explicitly for the importance of information technology in organizational life" (ix).

This framework shows how "institutions influence and are influenced by enacted information technologies and predominant organizational forms" (89). The author distinguishes objective technology (the Internet, hardware, software, etc.) from enacted technology ("the perception of users as well as designs and uses in particular settings") (10). Organizational forms refer to different types of organization, with the author focusing on bureaucracy and inter-organizational networks in their analysis. Finally, institutional arrangements "include the bureaucratic and network forms of organization and … institutional logics" (98). Therefore, the author concludes that the outcomes of technology enactment are a result of this complex interflow of relations and logics, and as such are multiple and unpredictable.

As can be seen in these two approaches to the relationship between human agency and technology, technology as an object in itself is very different from technology in use. Technology in use derives its meaning, implication and effects from contextual factors, such that it both constitutes and reflects back the world around it. Seen from this perspective, AI by itself is an 'objective technology', but once it is used it reflexively influences and is influenced by human agency and various institutional arrangements/organizational forms, leading to unforeseen consequences.

Gender bias, according to the European Institute for Gender Equality (2023), refers to "prejudiced actions or thoughts based on the gender-based perception that women are not equal to men in rights and dignity". This, therefore, constitutes the underlying mechanism for how gender biases influence and are influenced by technologies. While AI itself might be seen as a neutral objective technology, it is imbued with new meanings and implications through its use in specific contexts by humans (Fountain's 'enacted technology' or Orlikowksi's 'social construction' of technology'). As gender biases are implicit in our society and culture, they become part of the 'contextual factors' which influence the use of and understanding of AI technologies, which in turn become themselves embedded with the same biases.

This definition, as well as many other studies, demonstrates how this bias is often expressed through language. For example, research by Menegatti and Rubini (2017:1–2) suggests that asymmetrical power relations between the genders are expressed through stereotypes associated with everyday lexical choices (where traits such as 'nice, caring, and generous' are used to describe females while 'efficient,

agentic, and assertive' are used to describe men). However, they also point out that the idea of the male as the 'prototypical human being' is encoded in the structure of many languages, for example where 'chairman' refers to both sexes in English.

Another example is the AI service 'Genderify', launched in 2020, which uses a person's name, username and email address to identify their gender (Vincent 2020). Names beginning with 'Dr' seemed to consistently be treated as male, as "Dr Meghan Smith" was identified as having a 75.90% likelihood of belonging to a male. Elsewhere recent research describes automated robots which were trained on large datasets and standard models, but were found to exhibit strongly stereotypical and biased behaviour in terms of gender and race (Hundt et al. 2022).

Gender bias can also present itself through stereotypical imagery, with Schwemmer et al. (2020:1) asserting that "bias in the visual representation of women and men has been endemic throughout the history of media, journalism, and advertising". Studies conducted on gender stereotypes in science education resources (Kerkhoven et al. 2016:1), school textbooks (Amini and Birjandi 2012:138) and commercial films (Jang et al. 2019:198) all reveal the gendered representation of men and women in public images. However, this phenomenon, and particularly stereotypes embedded in digital or online imagery remains understudied (Singh et al. 2020:1282).

Here it is important to note that this study is concerned with 'gender' bias. This paper will follow the World Health Organization in defining gender as "the socially constructed characteristics of women and men—such as norms, roles and relationships of and between groups of women and men", which is distinct from 'sex': "the different biological and physiological characteristics of males and females, such as reproductive organs, chromosomes, hormones, etc." (in Council of Europe, 2023).

The above definition of 'gender' could itself be said to reinforce gender binaries—the idea that gender can be divided into two neat categories of 'male' and 'female', rather than representing the diverse spectrum of gender identities existent in society. However, as the purpose of this study is to summarize current research in the gender bias field, the majority of which assumes a binary definition of gender, this study will tentatively retain the above definition.

2.1 Related work

According to John McCarthy, a professor at Stanford University who first coined the term, AI is "the science and engineering of making intelligent machines, especially intelligent computer programs" (2007: 2). These programs are run on algorithms which are designed to make decisions or create solutions to a particular problem (West and Allen 2018). Algorithms are often seen as fairer or more neutral than humans in terms of decision-making (Gutiérrez, 2021:441).

However, as these systems are created by humans and fed with data based on the human experience, they inevitably also reflect inherent human biases. For example, in Caroline Perez's (2019:25) influential work on the gender data gap she explains how "we have positioned women as a deviation from standard humanity and this is why they have been allowed to become invisible". Thus, algorithmic bias can be generally defined as "the application of an algorithm that compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability, or sexual orientation and amplifies inequities in…systems" (Igoe 2021). Friedman and Nissenbaum's 1996 work on bias in computer systems points to three types of bias; pre-existing bias (emerging from societal attitudes and practices), technical bias (due to technological constraints) and emergent bias (which arises as the computer system is used). However, AI bias is an extremely complex topic, covering different forms of bias and notions of fairness (Bernagozzi et al. 2021:53).

Currently, there are two streams of literature that address gender bias. The first stream of literature focuses on pointing out the amplification of gender bias (often meaning discrimination against women) inherent in many technologies, such as in audio-visual data (Gutiérrez 2021), online language translators (Bernagozzi 2021) and recruiting tools (Dastin 2022).

The second stream of literature goes beyond exploring the existence of gender bias in technology, and additionally explores methods for mitigating this bias. This includes studies on how to reduce gender bias during the resume screening process (Deshpande et al. 2020), in machine learning models (Feldman and Peake 2021) and facial recognition systems (Dhar 2020). This stream includes both research on how to mitigate the effects of bias amplification which can be seen in AI, as well as studies which specifically aim to harness AI in order to reduce gender bias in technologies. Therefore, one essential aspect of our framework is to examine gender bias perpetuation and gender bias mitigation.

3 Analytical framework

Discussions on gender bias often naturally fall into two categories: studies which explore or attempt to measure gender bias in AI techniques (Stanovsky et al. 2019; Sheng et al. 2019;), and those which focus more on how to mitigate gender bias itself (Stafanovičs et al. 2020; Deshpande et al. 2020; Domnich and Anbarjafari 2021). This distinction has been noted by authors such as Blodgett et al. (2020), whose paper critically reviewing papers on bias in NLP notes that these studies either "proposed quantitative techniques for

Table 1 Classification of case studies based on content

AI data source	Bias perpetuation	Bias mitigation
Text	Prates et al. (2018). Assessing gender bias in machine translation: a case study with google translate	Bolukbasi T (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings
Image	Buolamwini and Gebru (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification	Wang T et al. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations

measuring or mitigating 'bias'" (1), or the Brookings Institute research framework for 'algorithmic hygiene' which includes identifying sources of bias and then forwarding recommendations on how to mitigate them (Lee et al. 2019). Of course, many of the studies which focus on mitigation techniques also implicitly or explicitly include descriptions or measurements of the gender bias issue they are attempting to resolve.

3.1 Defining bias perpetuation and mitigation

3.1.1 Bias perpetuation

Algorithms "don't simply reflect back social inequities but may ultimately exacerbate them" (Igoe 2021). Busuioc notes how algorithmic tools can "get caught in negative feedback loops" which then becomes the base for future predictions—all exacerbated if the initial data fed into the machine was itself biased (2021:826). Studies on the use of AI have discovered gender bias in the outcomes of algorithm application, from natural language processing techniques which perpetuate gender stereotypes (Kay et al. 2015) to facial recognition software which is much more accurate on male faces than female ones (Domnich and Anbarjafari 2021). Thus, the topic of 'algorithmic bias' in AI systems has become an important issue in the public sector, with concerns about how algorithms could systematize bias as well as questions around oversight and accountability mechanisms for these technologies (Alon-Barkat and Busuioc 2022: 5).

3.1.2 Bias mitigation

Bias mitigation involves "proactively addressing factors which contribute to bias" (Lee et al. 2019). In terms of algorithms, bias mitigation is often strongly associated with the concept of 'fairness'. For example, several researchers came together in 2018 to create the AI Fairness 360 (AIF360), a toolkit which provides a framework against which researchers can evaluate algorithms. This includes "bias mitigation algorithms" which can "improve the fairness metrics by modifying the training data, the learning algorithm, or the predictions" during the pre-processing, in-processing, and post-processing stages (Bellamy et al. 2019:7).

Mitigation techniques can take several forms, such as rebalancing data, regularization, adversarial learning etc. (Tomalin et al. 2021:70). Several studies look at mitigation in terms of deliberately designing 'fairer' algorithms (Salazar et al. 2021) As this study looks at algorithmic bias in different categories of AI technologies, it will focus on looking at the bias mitigation techniques which are being used to target bias in these technologies themselves.

3.2 Defining types of AI technologies

Corea (2019:26) created an AI Knowledge Map (AIKM) in order to represent and classify AI technologies. This categorizes AI technologies based on two macro-groups; AI paradigms and AI Problem Domains. Corea suggests six AI paradigms: logic-based tools, knowledge-based tools, probabilistic methods, machine learning, embodied intelligence and search and optimization. These are plotted against the problems AI has been used to solve, here divided into five clusters related to reasoning, knowledge, planning, communication and (sensory) perception.

For this paper, we are more interested in the way Corea groups technologies under the statistical AI paradigm, including Computer Vision, Neural Networks and Natural Language Processing. The scope of this investigation will be within this set of statistical methods, and as is common practice in the computer science field, further classified based on the data source being either image or text (Table 1). Therefore, the focus will largely be on the areas of Computer Vision and Natural Language Processing.

3.3 Method and case analysis

The framework of this study looks at bias perpetuation and mitigation across different AI technologies. Therefore, at first a general literature review was conducted using keywords to collate a range of materials relating to bias perpetuation and mitigation in several AI technologies.

On the basis of a further initial analysis of relevance as well as whether or not the reports could be paired into perpetuation/mitigation pairs for different technologies, these papers were further narrowed down to the final four.

Then case information was gathered through an information selection process. Information was included based on strict criteria which looked to find the following.

- Location and timing of the study
- Motivation/purposes for such AI adoption
- Process of implementation or creation
- Gender bias issue
- Host agencies (agencies that adopted the technologies)
- Relevant stakeholders (government non-profits, universities/schools, etc.)
- Effects/influence on gender bias issue
- Relevant regulations/policies

This information was then written up into the case studies below.

4 Cases of AI and gender biases/mitigation

4.1 Bias perpetuation in NLP

In 2018, a group of researchers at the Federal University of Rio Grande do Sul in Brazil decided to test the existence of gender bias in AI, specifically in automated translation (Prates et al. 2020).

In the experiment, they ran the sentence constructions in the form 'He/She is a [job position]' (for example, 'He/She is an engineer') from English into twelve languages which are gender neutral using Google Translate. A gender-neutral language is one in which there are not separate male and female pronouns, for example, whereas in English the male pronoun 'He' and the female pronoun 'She' are separate words, in Hungarian the pronoun 'ő' can represent both 'he' and 'she'. The twelve languages they chose were Malay, Estonian, Finnish, Hungarian, Armenian, Bengali, Japanese, Turkish, Yoruba, Basque, Swahili and Chinese.

They then selected job positions from a list issued by the U.S. Bureau of Labour Statistics (BLS), which also gives the percentage of women participation in these occupations. The researchers ran the 'He/She is a [job position]' sentence through Google Translate, noting how often the translation of the gender-neutral pronoun came out as 'He' or 'She'. They expected that this translation tool would reflect the inequalities in society, and therefore inevitably display some bias in assuming certain pronouns for certain jobs. For example, at the time of the research, translating various sentences using the construction 'He/She is a [job position]' with the gender-neutral pronoun 'ő' from Hungarian to English gave stereotyped results, such as 'She's a nurse', 'He is a scientist', 'He is an engineer' (where 'He's a nurse', 'She is a scientist' or 'She is an engineer' would have been equally correct).

The authors found that machine translation is strongly biased towards male defaults, especially for fields such as STEM which are typically thought of as weighted towards one gender. These results also did not reflect real-world statistics on gender ratios in this field. For example, 39.8% of women work in the category of 'management', but sentences were translated with a female pronoun only 11.232% of the time (66.667% of the time as male, and 12.681% of the time neutrally). Overall, women made up 35.94% percent of the BLS occupations, but sentences were only translated with female pronouns 11.76% of the time, showing the translations do not reflect workplace demographics. These results did vary across language, as translations from Japanese and Chinese produced female pronouns only 0.196% and 1.865% of the time respectively, while Basque produced a majority of gender-neutral pronouns.

The authors also completed a similar subset of research using commonly used adjectives to describe human beings, including 'Happy' 'Sad' 'Shy' 'Polite' etc. This produced a more varied mixture of results, where words such as 'Shy', 'Attractive', 'Happy', 'Kind' and 'Ashamed' tended to be translated with female pronouns, while 'Arrogant', 'Cruel' and 'Guilty' tended towards male pronouns (with 'Guilty' in fact being exclusively translated with a male pronoun for all languages).

The authors noted that a few months after their research was published on Cornell University-based arXiv.org open repository, Google released a statement admitting the possibility of gender bias in the Google Translate system and revealing a new feature whereby when a sentence is translated from a language which has gender-neutral pronouns to one which has gendered pronouns, both male and female pronouns are presented as possible translations. For example, now translating the sentence 'ő egy orvos' from Hungarian to English presents two translations: 'she is a doctor' and 'he is a doctor'. A link to a separate page, which offers further information about gender-specific translations is also included, noting the current gender-specific translation options as well as stating that 'Gender-specific translations in more languages are coming soon'.

More recently, in 2021, Google Translate released the 'Translated Wikipedia Biographies' dataset, through which gender bias of machine translation can be measured, due to the high potential for translation errors—they state that datasets can reduce errors by 67% (Stella 2021).

The paper also gained traction in both scholarly and journalistic circles. The article has currently been cited over 180 times, including by authors studying bias in machine translation such as Font and Costa-Jussa (2019), who cite Prates' study as one which has detected the problem of gender bias in this technology. Later studies such as Savoldi et al. (2021) go on to extend Prates et al.'s work, noting the prevalence of evidence showing occupational stereotyping, while pointing

Table 2 Snapshot of Bolukbasi et al. (2016) results

Feminine	Tote; Ultrasound; Flirt; Divorce; Tearful; Modeling; Crafts; Browsing; Busy; Trimester
Masculine	Buddy; Command; Firepower; Game; Zeal; Guru; Yard; Youth; Firmly; Builder

out the importance of studying other relevant phenomena such as gendered associations of characteristics or psychological traits, as well as to go further than this by investigating if/whether these can cause harms and if so, how.

4.2 Bias mitigation in NLP

In 2016, a group of researchers from Boston University and Microsoft's Research Lab in New England, USA, came together to propose a methodology for removing gender bias from word embeddings—a natural language processing task which captures semantic associations between words in a text (Bolukbasi et al. 2016). A word embedding represents each word in text data as a 'word vector', which is a mathematical representation of the meaning of the word by mapping it in space (Alizadeh 2021).

This provides two sets of information about word meanings in a text. Firstly, vectors which are closer together represent words which have similar meanings. Secondly, comparing different vectors can represent semantic relationships between words, enabling the input of 'man is to king as woman is to x' to find x = 'queen' (2016:1). The researchers note that there is much research on word embeddings themselves; however, little attention is paid to the inherent sexism captured by word embeddings, which will predict the answer to 'man is to computer programmer as woman is to x' as 'x = homemaker'. As word embeddings are widely used as a basic feature in NLP, their use has the potential to amplify gender bias in systems.

In this paper, the researchers analysed the 'word2vec' embedding. This is a popularly used embedding which uses neural network methods to learn embeddings from data sets (TensorFlow 2022). The embedding is trained on a 3 million word-large English language Google News corpus and the resulting embedding is referred to by the researchers as 'w2vNEWS'. The aim of the study is to first demonstrate the biases contained in word embeddings, and then to create a debiasing algorithm to "remove gender pair associations for gender-neutral words".

'Gender neutral words' are words which have no specific gender association and these are contrasted with gender-specific words which explicitly include a gendered reference. For example, 'daughter' 'lady' and 'queen' are examples of gender-specific words as they explicitly refer to the female gender. However, 'rule' 'game' 'nurse' and 'homemaker' are all examples of gender-neutral words—words which do not refer to one gender or the other. Yet, despite this, 'gender-neutral words' often are semantically correlated to a certain gender. Thus, the authors found that certain words such as 'cocky', 'genius' and 'tactical' were all associated with the male, while 'tanning', 'beautiful' and 'busy' were all vocabulary associated with the female (Table 2). The table below summarizes selected words which the researchers found had a gendered association:

The debiasing algorithm developed by the researchers aimed to remove the gender pair associations for all these 'gender neutral' words, while retaining the function of word embedding in mapping useful relationships and associations between words. The algorithm involves two steps, the first step identifies the subspace that shows the gendered bias. The second step either 'neutralizes and equalizes' (gets entirely rid of the gendered connotations of gender-neutral words and then ensures they are equidistant from all other words in the set) or 'softens' the bias (maintaining certain useful distinctions between words in a set—for example where a word has more than one meaning). They then evaluated the algorithm through generating word pairs comparable to 'she-he' (for example 'he' is to 'doctor as 'she' is to 'x', where the algorithm must determine the value of x) before asking crowd workers to rate whether these pairs reflected gender stereotypes.

While the initial embedding was found to represent stereotypes 19% of the time, the new debasing algorithm reduced this percentage to 6%. For example, they noted that the original embedding would find the x in 'he is to doctor as she is to X' as 'nurse'; however, the new embedding found 'x = physician'. Despite this, the algorithm still preserved appropriate analogies, such as 'she is to 'ovarian cancer'' as 'he is to 'prostate cancer''.

The authors noted that to entirely solve this problem "one should attempt to debias society rather than word embedding"; however, they note that their algorithm at the very least will not amplify bias (8). This research has been cited over 1000 times in studies on AI ethics, bias in machine learning and papers on bias mitigation. It has also been cited by popular news sites such as Forbes (Roselli et al. 2019) and The Conversation (Zou 2016) as well as on academic sites such as MIT Technology Review (2016). The code itself is available on GitHub for users to download themselves and debias their own text data[1].

[1] See here: https://github.com/tolga-b/debiaswe

4.3 Bias in digital images

This study, carried out by Joy Buolamwini of MIT and Timnit Gebru from Microsoft Research (2018), begins by pointing out how facial recognition tools are starting to be used in a public administration capacity, including in the criminal justice system. Despite this, there is a lack of studies which look at how to create fairer algorithms and mitigate biases in this area.

There are several established benchmarks used to test computer vision programs, a benchmark being a standardized set of data against which algorithms are tested to determine their accuracy. In terms of computer vision, these benchmarks are a group of images showing faces of those across the gender and racial spectrum. However, the researchers found that current benchmarks overrepresented lighter skinned and male individuals, while underrepresenting darker skinned individuals. Therefore, they first created their own benchmark—the Pilot Parliaments Benchmark—where the set of individual faces were selected to cover an equal mix of darker and lighter skinned, male and female individuals.

They then used this data to evaluate three commercial data classifiers, including Microsoft's Cognitive Services Face API, IBM's Watson Visual Recognition API and Face^{++}, a company headquartered in China providing technology which has previously been integrated into Lenovo computers.

Their results revealed intersectional errors in the software, which was routinely less accurate for women than men, and for darker skinned individuals than lighter skinned individuals. The classifiers wrongly recognized female faces more often than men's, with error rates between female and male classification ranging from 8.1% to 20.6%. For the Face^{++} classifier, false positives (where the algorithm wrongly identified the gender of the face) were returned 13 times more often for women than men. The authors also conducted an intersectional analysis, showing that darker skinned females are most likely to be misclassified, returning 61.0% to 72.4.1% of the classification error despite making up only 21.3% of the benchmark.

The authors conclude by calling for improvement of classification of darker skinned individuals as well as closing the error gap between male and female classification. They point to more inclusive face datasets and algorithmic evaluation as areas for future research, including improving accountability and transparency in benchmark datasets and algorithmic performance.

This article has been cited extensively (over 2600 times) and 'Gender Shades' now has its own web page, allowing the user to explore the results of the paper and related issues[2]. IBM issued a direct response to the study, thanking them for contributing to the conversation around data ethics and AI. They state that they have been working to increase the accuracy of their facial analysis software, and now use "different training data and different recognition capabilities" than the software used in the study[3]. The company then ran the data through benchmark images very similar to those from the study and showed the returning of a much lower rate of errors (although the greatest errors were still with darker skinned female images). They then described the various algorithms they are developing to detect, rate and correct bias both in data and models.

Microsoft also sent a response to the lead author, reinforcing their support for fairness in AI technologies and stating that they too had taken steps to improve the accuracy of their technology[4]. Face^{++} did not respond to the results.

The Gender Shades website also links to the Algorithmic Justice League, an organization looking to reduce bias in coding, while also making its Pilot Parliaments Benchmark available to request. Facial recognition is used for two types of tasks, verification (comparing the selected image to one other—for example the iPhone's FaceID) and identification (identify whether the selected image corresponds to another image in a gallery—used for finding missing persons or criminal matches) (MIT Media Lab-a, 2018).

The authors emphasize that false positives could threaten civil liberties of individuals (Buolamwini and Gebru 2018:2). They sort potential harms from biased algorithmic decision-making into three main categories, including 'loss of opportunity' (in hiring, housing, education etc.), 'economic loss' (credit etc.) and 'social stigmatization' (stereotype reinforcement, dignitary harms, etc.), emphasizing that these are both individual and collective societal harms (MIT Media Lab-b, 2018).

4.4 Gender bias mitigation in digital images

Researchers from the University of Virginia, University of California Los Angeles and the Allen Institute for Artificial Intelligence came together in 2019 to explore the issue of gender bias in image representation (Wang et al. 2019). Their study begins by pointing out how facial recognition systems often amplify biases based on protected characteristics such as race or gender, and how this can have real-world consequences, for example autonomous vehicle systems being unable to recognize certain groups of people.

They begin by studying bias amplification through the COCO dataset for recognizing objects and the imSitu dataset

[2] See here: http://gendershades.org/

[3] See response: http://gendershades.org/docs/ibm.pdf

[4] See response here: http://gendershades.org/docs/msft.pdf

for recognizing actions. The COCO dataset (Microsoft Common Objects in Context) is an image dataset which can be used to train machine learning models, containing over 328,000 annotated images of humans and every-day situations (datagen, 2022). The imSitu dataset contains images describing situations along with annotations describing the situations, which can also be used to train algorithms on situation recognition[5].

They propose a new definition for measuring bias amplification, where instead of comparing the training data and model predictions, they compare "the predictability of gender from ground truth labels (dataset leakage…) and model predictions (model leakage…)" (2019:5310). Ground truth labels are those labels assigned to the data by human workers—that is to say they are accurate representations of the data. Model predictions are those made by the model (algorithm) itself, and thus comparing these two makes it possible to test the accuracy of the modelling. Using this method, they find that even models which are not programmed for predicting gender will still amplify gender bias.

They hypothesize that models may perpetuate biases because there are gender-related features in the image which are not labelled by the computer program, but may still be taken into account when predicting gender—this is called 'data leakage' in this paper. For example, they give the example of a dataset with an equal number of women and men shown cooking. This in itself does not amplify bias, but if there is a child in the image, and children are often shown more with women than men across all images, then the model may associate 'children' with 'cooking', and therefore overall women could be labelled as 'cooking' more than men still. Model leakage then referred to how much the model's predictions were able to identify protected characteristics (here gender).

The researchers adopted the method of 'adversarial debiasing' in order to mitigate this effect. This could preserve useful information, while removing gender correlated features in the images. Sometimes this involves eliminating the face, or even gender-associated clothing, while retaining information needed to recognize actions or objects. Their proposed algorithm aims to "build representations from which protected attributes can not be predicted" (5315).

Quantitatively, the algorithm was able to reduce model leakage by 53% for COCO and 67% for imSitu. Then, comparing their method with another debiasing algorithm (RBA), they show that the authors' methods are much more effective at reducing bias amplification.

Overall, they conclude that balanced datasets are not enough to prevent encoded bias in computer vision, and instead support the idea of removing features associated with a protected variable (such as gender) from images. Their work has been cited over 160 times. Their code is available online, and as well as this, they have created a demo page where users can upload their own image and apply the adversarially trained neural network to obscure gender information.

5 Discussion and implications

Existing frameworks relating to investigating AI bias can be split into two general streams. One stream includes frameworks which address a very specific area of AI bias, such as Blodgett et al. 2020 framework which addresses the area of bias in NLP, surveying related research and dividing it in terms of specific types NLP bias, or Gutiérrez's (2021) study which specifically investigated bias related to audio-visual technologies. The second stream includes frameworks which attempt to cover the entire spectrum of AI bias's emergence and mitigation. For example, Ferrer et al. 2021 discussion of bias in AI applies a cross-disciplinary lens to this topic, covering bias in the modelling, training and usage of AI, and as such attempting to capture a picture of AI across its entire ecosystem.

The framework put forward in this paper combines these two approaches, in that it is both flexible enough to encompass a large range of AI-related biases, while also dividing these into useful categories. One recurring issue in AI research is that it is often said to replicate the 'black box' of algorithms, being difficult to understand from outside the field. This framework thus intentionally simplifies the field of AI biases into an accessible two-by-two grid based on the AI data source used, whether this be image or text, as well as whether it is an example of bias perpetuation or mitigation. In this way, it is flexible enough to capture the breadth of existent AI bias, while also presenting these in easily recognizable categories.

5.1 Implications

5.1.1 Text: bias perpetuation

The study by Prates et al. stimulated major interest, both from the industry and academic circles, as well as being reported in the mainstream media and being the subject of follow-up research. The study reveals how machine translation, although used as a neutral tool by many, can unwittingly perpetuate gender bias through its usage (Table 3). On one hand, it shows the importance of training data—as Google Translate is trained on existing data, it replicates the biases found in this data. Interestingly, it also emphasizes the differences in machine translation biases across languages, showing that any changes to technology would have to take

[5] See: http://imsitu.org/

Table 3 Implications of each case study

AI data source	Bias perpetuation	Bias mitigation
Text	Address data biases Address algorithmic bias (Vanmassenhove 2019)	Introduction of neutral words Mitigation of word2vec embedding Use debiased embeddings Machine learning may not be able to eliminate gender bias in society, but it can try to at least not amplify it
Image	Address training data biases Create new benchmark for digital images to reduce data biases (benchmark = databases of images on which facial recognition software is trained)	Even unbiased data sets (benchmarks) can lead to bias in recognition These results can be extended to other sources of bias

into account these differences as well. Vanmassenhove (in Diño, 2019) has also stated that in terms of attempts to mitigate gender bias in machine translation, different solutions need to be considered for different languages—"It is not translating from [the aforementioned] languages into English that is problematic, but the other way around. Different languages have different ways of expressing gender and it is important to realize that there won't be one solution that fits all".

However, she has pointed out that the larger problem with neural networks is that they "do not just reflect controversial societal asymmetries but 'exaggerate' them". While de-biasing training data has been suggested as a mitigation technique, she reflects that biases go beyond just gender—scrubbing gender bias may solve one issue, but other biases (based on race, age, etc.) will still remain.

Aside from this, the study also shows the importance of pointing out these issues and their real-world consequences; the study was able to stimulate Google to redesign its systems to mitigate the original bias which occurred. In this way, it is important to investigate and hold accountable different algorithmic results, which in turn may prompt companies to more effectively and actively continually attempt to remove bias.

5.1.2 Text: bias mitigation

The issue of word embeddings is often cited in relation to gender bias, as latent gendered word associations can be clearly mapped using this technology. The study by Bolukbasi et al. reminds us of the importance of retaining the utility of algorithms and AI technologies while aiming to minimize the level of bias in them. Therefore, the researchers did not aim to eliminate bias in their results, but instead aimed to ensure that the algorithm could still capture useful semantic relationships between words in a text, while avoiding the replication of gender bias.

Thus, although it was called a 'debiasing algorithm', the researchers described its function as 'softening' the bias. Therefore, although they assert the importance of 'debiasing society' as a prerequisite for 'debiasing' AI technology, they do still point to the value of technology in minimizing existing discrimination.

5.1.3 Images: bias perpetuation

A key takeaway from the study by Buolamwini and Gebru (2018) is that bias is perpetuated not just by training data or modelling systems, but also by the method of evaluating these systems. They find that the 'benchmarks', or sets of images used generally for testing visual data themselves reflect bias, driving the researchers to develop their own new benchmarks. Therefore, this points to the importance of accountability in algorithmic development through developing accurate ways to assess not only the bias inherent in the algorithms themselves, but also in the tests used to assess their accuracy and bias perpetuation.

The Gender Shades project also shows the real-world consequences of discrimination for people who are misidentified by these decisions, and therefore works as an interesting case study in how rigorous scientific research can be combined into a wider cross-disciplinary movement for racial justice and made understandable to a wider audience. It is also a study which demonstrates why research on AI bias is relevant. A critical normative question which remains unaddressed in many studies, is what exactly counts as 'gender bias', 'bias' or 'discrimination'. Blodgett et al.'s 2020 paper critically surveyed all of the NLP related papers which looked at measuring or mitigating bias and picked out several areas for improvement, mainly focused around issues of motivation. They noted that papers lacked normative reasoning for why the study was undertaken, as well as lacking a normative understanding or definition of bias.

This led to an additional problem of papers not explicitly stating *why* or *how* biases could harm different groups. In their conclusion they proposed that going forwards there should be more normative efforts to explain what 'biases' are being investigated, who they harm and why, as well as more cross-discipline studies which include relevant literature on language and societal hierarchies which illuminate

ingrained power relations. Although several of the papers mentioned in this report—perhaps most explicitly Bolukbasi et al's. Gender Shades project—go some way in detailing the real-world impact of the algorithmic bias found, there is still a lack of cross-disciplinary research which goes beyond pointing out errors or biases on a technological level, but also links this to a wider socio-cultural context.

5.1.4 Images: bias mitigation

Finally, the research by Wang et al. (2019) emphasizes the fact that policy guidelines should recognize gender biases reinforced through applying AI in text datasets in terms of *data biases* and *algorithmic biases*. For instance, while Prates et al. (2019) illustrates data biases which male or female nouns are associated with particular professions, Vanmassenhove (2020) reveals an algorithmic bias on the GNMT systems when findings overgenerate male or female nouns even taking into account the data bias. The study by Wang et al. also shows how balanced datasets are not enough to reduce gender biases, and that algorithms can and do actively amplify what they term 'protected characteristics' such as racial or gendered characteristics.

However, it also illustrates the importance of combatting bias in a holistic manner. The research highlights that much of the image bias was not necessarily due to aspects present in the images, but because of gendered associations between different concepts. Therefore, any attempt to eliminate bias should not focus separately on different aspects of the bias perpetuation, but must consider the whole process of bias perpetuation and mitigation to ensure that the problem is tackled at its source, rather than just tackling certain symptoms.

Os Keyes, a PhD Candidate at the University of Washington's Department of Human Centred Design & Engineering, brings up a similar view to Orlikowski when they point out that "Far too often, work in this area looks for examples of explicit gender bias, operating from the implicit assumption that absent that [bias], technology (and society) are 'neutral'", instead insisting that what is more important is 'how we frame "gender bias': what we look for, where we look for it, and how we draw connections between different types of and sources of oppression" (in Bryson et al. 2020).

The underlying premise behind all of the above is that AI and gender bias research must embrace a cross-disciplinary and accessible approach. This means linking ideas from gender studies, sociology, linguistics, technology, ethics and public administration, in order to form a holistic approach to mitigating algorithmic gender bias, a dialogue which Noble (2018:13) believes is necessary "before blunt artificial intelligence decision-making trumps nuanced human decision making".

Finally, we must recognize that efforts are part of an overall process. It is impossible to eliminate all gender bias in technology, because gender bias is inherent in all areas of culture and society. Therefore, it is essential to understand this concept, and see important efforts as regulatory ones—to expect, point out and regulate gender bias which becomes apparent in technologies. As part of an intersectional approach, this will inevitably not be limited to just gender, but other biases as well.

5.2 Policy recommendations

In Noble's work on the 'Algorithms of Oppression' (2018:1), she posits that "artificial intelligence will become a major human rights issue in the twenty-first century". In this way, there have already been attempts by national and international institutions to begin creating policies and frameworks to identify and mitigate these biases. In 2019 the independent High-Level Expert Group on Artificial Intelligence, set up by the European Commission, produced a report entitled 'Ethics Guidelines for Trustworthy AI'. The paper proposes "equality, non-discrimination and solidarity" as a fundamental right, calling to ensure that systems do not generate unfairly biased outputs, including using inclusive data which represents different population groups. The European Commission also aims to introduce a legal framework for AI, in compliance with the E.U. Charter of Fundamental Rights, aimed at defining responsibilities of users and providers (Di Noia et al. 2022).

This charter aims to balance the innovation and positives of new AI technologies with the basic rights of EU citizens as well as bringing them in line with EU values. It proposes a 'risk-based' approach, by aiming to prohibit artificial intelligence practices which are not in accordance with EU laws or values. Although most of the case studies in this paper shine a light on specific aspects of AI bias, the legal framework aims to reduce especially the actual biases in decision-making which could affect its citizens.

A separate report by UNESCO (2020) on AI and gender equality suggests a range of practices for integrating gender equality into AI principles, including proactive mitigation, making the invisible visible and understanding AI as a potentially empowering tool for girls and women. Many of the case studies in this paper point out that bias is inherent in society and thus it is inherent in AI as well. The UNESCO recommendations accept this premise, but still promote the importance of "shift[ing] the narrative of AI as something 'external' or technologically deterministic, to something 'human' that is not happening to us but is created, directed, controlled by human beings and reflective of society". Thus, it is not a question of either/or as to whether to first change society or change AI, both must be achieved in tandem.

A similar sentiment is echoed in a review on Artificial Intelligence and Public Standards by the Committee on Standards in Public Life, an independent body advising the UK government. Their report (2022) concludes that the existing 'Seven Principles of Public Life' (selflessness, integrity, objectivity, accountability, openness, honesty and leadership) should be upheld as a guide in how to integrate AI technologies into public life. Understanding that AI may have wide-ranging and unexpected effects, the review proposes a general outline of how the Seven Principles can be translated into practice for the use of AI (16). Overall, it is clear that current policy recommendations for the regulation of AI focus on overarching principles and guidelines, reflecting the ongoing and expanding range of issues which may need to be addressed in future.

6 Conclusions

Our study not only brings awareness to the potential gender biases caused by the current AI technologies, but also highlights the potential of mitigation of gender biases from the AI technologies. As shown by UN and EU attempts to propose new policies and principles to regulate the potential effects of AI on gender equality, policy makers are yet to reach a consensus on how to balance AI's potential for empowering women with the possible detrimental effect it could also have.

It is clear that greater collaboration is necessary across different sectors affected by these issues. Tannenbaum et al. (2019: 137) write on "the potential for sex and gender analysis to foster scientific discovery, improve experimental efficiency and enable social equality", positing that "integrating sex and gender analysis into the design of research, where relevant, can lead to discovery and improved research methodology". They propose a framework to ensure issues of sex and gender are considered in scientific and engineering research, consisting of coordinated policies between funding agencies, universities and peer-reviewed journals, combined with "development of methods of sex and gender analysis" by evaluators and researchers, as well as "greater rigor, reproducibility, inclusion and transparency" in research in general (143).

As described by the case studies in this paper, the barrier to access in fully understanding technological workings and implications is high, limiting opportunities for comprehensive studies covering both the technological, social and policy effects of such tools. However, as described above by Tannenbaum et al. the integration of research from the fields of science and technology, gender studies and public policy is necessary in order to reach a common understanding of the broad benefits and challenges in introducing complicated technologies to modern life.

For example, Crenshaw's (1989) theory of intersectionality, exploring how the treatment of women can vary depending on the interaction of their various identities (for example race, disability etc.), is clearly relevant in helping to describe how considering women as a homogenous group can itself obscure the huge variation in treatment between different groups. This raises the question of how AI gender bias can address intersectional biases which arise.

Furthermore, with a growing debate on the concept of gender itself and the proliferation of research into variations in gender identity, perhaps consideration should be applied to the binary concept of gender often applied in related research, which can again hide or ignore the experience of those who do not conform to traditional gendered expectations or labels.

Finally, from the case studies analysed above, it is clear that discussions around accountability and ethical responsibilities in AI technology must take centre stage going forwards. Although this paper has focused specifically on gender bias, this is just one of many other areas of discrimination evident in society which has the potential to be filtered or reflected through the prism of algorithms. As the concept of intersectionality makes clear, the mitigation of one type of bias does not solve the issue of inequality—a holistic effort to mitigate interacting biases is the only way to effectively improve the fairness of outcomes for all.

Acknowledgements The authors would also like to acknowledge the advice and suggestions given by Karen Lu, Graduate Assistant, Communications & Multimedia Laboratory, National Taiwan University, Taipei, Taiwan (R.O.C).

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by SOC and HL. KL contributed to the creations of the analytical framework. The first draft of the manuscript was written by SOC and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding National Science Council, MOST110CD804-1, Helen K. Liu.

Availability of data and materials Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

Agarwal PK (2018) Public administration challenges in the world of AI and bots. Public Adm Rev 78(6):917–921

Alizadeh K (2021) Word Vectors and Word Meanings. Medium. https://towardsdatascience.com/word-vectors-and-word-meaning-90493d13af76. Accessed 18 Jan 2023

Alon-Barkat S, Busuioc M (2022) Human-AI interactions in public sector decision-making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice. J Pub Adm Res Theory 33(1):153–169

Amini M, Birjandi P (2012) Gender bias in the Iranian High School EFL Textbooks. Engl Lang Teach 5(2):134–147

Bellamy RK, Dey K, Hind M et al (2019) AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM J Res Dev 63(4/5):4:1-4:15

Bernagozzi M, Srivastava B, Rossi F, Usmani S (2021) Gender bias in online language translators: visualization, human perception, and bias/accuracy tradeoffs. IEEE Internet Comput 25(5):53–63

Blodgett SL, Barocas S, Daumé III H, Wallach H (2020) Language (Technology) is Power: A Critical Survey of "Bias" in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Adv Neural Inf Process Syst 29:4349–4357

Bryson J, Etlinger S, Keyes O, Rankin JL (2020) Gender Bias in Technology: How Far Have We Come and What Comes Next? CIGI. https://www.cigionline.org/articles/gender-bias-technology-how-far-have-we-come-and-what-comes-next/?utm_medium=social&utm_source=twitter. Accessed 17 Jan 2023

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. Conf Fairness Account Transpar 81:77–91

Busuioc M (2021) Accountable artificial intelligence: Holding algorithms to account. Public Adm Rev 81(5):825–836

Corea F (2019) An introduction to data. Springer, Cham

Council of Europe (2023) Sex and gender. https://www.coe.int/en/web/gender-matters/sex-and-gender. Accessed 17 Jan 2023

Crenshaw K (1989) Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum, 1989(1), Article 8

Dastin J (2022) Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. In Martin K (ed) Ethics of Data and Analytics, 1st edn. Auerbach Publications, pp. 296–299

Datagen (2022) MS COCO Dataset: Using it in Your Computer Vision Projects. https://datagen.tech/guides/image-datasets/ms-coco-dataset-using-it-in-your-computer-vision-projects/. Accessed 17 Jan 2023

Deshpande KV, Pan S, Foulds JR (2020 July). Mitigating demographic Bias in AI-based resume filtering. In Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (pp. 268–275). https://doi.org/10.1145/3386392.3399569

Dhar P, Gleason J, Souri H, Castillo CD, Chellappa R (2020) Towards gender-neutral face descriptors for mitigating bias in face recognition. arXiv preprint arXiv:2006.07845

Di Noia T, Tintarev N, Fatourou P, Schedl M (2022) Recommender systems under European AI regulations. Commun ACM 65(4):69–73

Diño G (2019) He Said, She Said: Addressing Gender in Neural Machine Translation. Slator. https://slator.com/he-said-she-said-addressing-gender-in-neural-machine-translation/. Accessed 17 Jan 2023

Domnich A, Anbarjafari G (2021) Responsible AI: Gender bias assessment in emotion recognition. arXiv preprint arXiv:2103.11436

Ethics guidelines for trustworthy AI. Publications Office. https://data.europa.eu/doi/https://doi.org/10.2759/177365. Accessed 17 Jan 2023

European Commission, Directorate-General for Communications Networks, Content and Technology (2019) Ethics guidelines for trustworthy AI, Publications Office. https://data.europa.eu/doi/10.2759/346720. Accessed 23 April 2023

European Institute for Gender Equality (2023) Gender Bias. https://eige.europa.eu/thesaurus/terms/1155. Accessed 3rd Feb 2023

Feeney MK, Fusi F (2021) A critical analysis of the study of gender and technology in government. Inform Polity 26(2):115–129

Feldman T, Peake A (2021) End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. arXiv preprint arXiv:2104.02532.

Ferrer X, van Nuenen T, Such JM, Coté M, Criado N (2021) Bias and discrimination in AI: a cross-disciplinary perspective. IEEE Technol Soc Mag 40(2):72–80

Filgueiras F (2022) New Pythias of public administration: ambiguity and choice in AI systems as challenges for governance. AI & Soc 37(4):1473–1486

Font JE, Costa-Jussa MR (2019) Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv preprint arXiv:1901.03116

Fountain JE (2004) Building the virtual state: Information technology and institutional change. Brookings Institution Press

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inform Syst (TOIS) 14(3):330–347

Gutierrez M (2021) New feminist studies in audiovisual industries | algorithmic gender bias and audiovisual data: a research agenda. Int J Commun 15:439–461

Hundt A, Agnew W, Zeng V, Kacianka S, Gombolay M (2022, June). Robots Enact Malignant Stereotypes. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 743–756)

Igoe, KJ (2021) Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It. Harvard T. H Chan School of Public Health. https://www.hsph.harvard.edu/ecpe/how-to-prevent-algorithmic-bias-in-health-care/. Accessed 17 Jan 2023

Jang JY, Lee S, Lee B (2019) Quantification of gender representation bias in commercial films based on image analysis. Proceed ACM on Human-Comput Interact 3:1–29

Kerkhoven AH, Russo P, Land-Zandstra AM, Saxena A, Rodenburg FJ (2016) Gender stereotypes in science education resources: A visual content analysis. PLoS ONE 11(11):e0165037. https://doi.org/10.1371/journal.pone.0165037

Lee, NT, Resnick P, Barton G (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institute: Washington, DC, USA. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/. Accessed 17 Jan 2023

McCarthy, J. (2007). What is artificial intelligence? http://jmc.stanford.edu/articles/whatisai/whatisai.pdf. Accessed 3 February 2023

Menegatti M, Rubini M (2017) Gender bias and sexism in language. In Oxford Research Encyclopedia of Communication. https://oxfordre.com/communication/view/https://doi.org/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-470

MIT Media Lab-a (2018) Gender Shades Project: Frequently Asked Questions. https://www.media.mit.edu/projects/gender-shades/

faq/#faq-after-a-face-is-detected-what-sort-of-recognition-tasks-can-be-done. Accessed 17 Jan 2023
MIT Media Lab-b (2018) Gender Shades Project: Why This Matters. https://www.media.mit.edu/projects/gender-shades/why-this-matters/. Accessed 17 Jan 2023
MIT Technology Review (2016). How Vector Space Mathematics Reveals the Hidden Sexism in Language. https://www.technologyreview.com/2016/07/27/158634/how-vector-space-mathematics-reveals-the-hidden-sexism-in-language/. Accessed 17 Jan 2023
Noble SU (2018) Algorithms of oppression. New York University Press
Orlikowski WJ (1992) The duality of technology: Rethinking the concept of technology in organizations. Organ Sci 3(3):398–427
Perez CC (2019) Invisible women: Data bias in a world designed for men. Abrams
Prates MO, Avelar PH, Lamb LC (2020) Assessing gender bias in machine translation: a case study with google translate. Neural Comput Appl 32(10):6363–6381
Roselli D, Matthews J, Talagala N (2019) Managing Bias In AI: What Should Businesses Do? Forbes. https://www.forbes.com/sites/cognitiveworld/2019/05/29/managing-bias-in-ai-what-should-businesses-do/?sh=10503ad21440. Accessed 17 Jan 2023
Salazar T, Santos MS, Araújo H, Abreu PH (2021) FAWOS: fairness-aware oversampling algorithm based on distributions of sensitive attributes. IEEE Access 9:81370–81379
Savoldi B, Gaido M, Bentivogli L, Negri M, Turchi M (2021) Gender bias in machine translation. Trans Assoc Comput Linguist 9:845–874
Schwemmer C, Knight C, Bello-Pardo ED, Oklobdzija S, Schoonvelde M, Lockhart JW (2020) Diagnosing gender bias in image recognition systems. Socius 6:1–17
Sheng E, Chang KW, Natarajan P, Peng N (2019) The woman worked as a babysitter: On biases in language generation. arXiv preprint arXiv:1909.01326
Singh VK, Chayko M, Inamdar R, Floegel D (2020) Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. J Am Soc Inf Sci 71(11):1281–1294
Stafanovičs A, Bergmanis T, Pinnis M (2020) Mitigating gender bias in machine translation with target gender annotations. arXiv preprint arXiv:2010.06203.
Stanovsky G, Smith NA, Zettlemoyer L (2019) Evaluating gender bias in machine translation. arXiv preprint arXiv:1906.00591.
Stella R (2021) A Dataset for Studying Gender Bias in Translation. Google AI Blog. https://ai.googleblog.com/2021/06/a-dataset-for-studying-gender-bias-in.html. Accessed 18 Jan 2023
Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L (2019) Sex and gender analysis improves science and engineering. Nature 575(7781):137–146
TensorFlow (2022) word2vec. https://www.tensorflow.org/tutorials/text/word2vec. Accessed 17 Jan 2023
Tomalin M, Byrne B, Concannon S, Saunders D, Ullmann S (2021) The practical ethics of bias reduction in machine translation: why domain adaptation is better than data debiasing. Ethics Inf Technol 23(3):419–433
UNESCO (2020, August) Artificial intelligence and gender equality: key findings of UNESCO's Global Dialogue (Document code: GEN/2020/AI/2 REV). https://unesdoc.unesco.org/ark:/48223/pf0000374174. Accessed 17 Jan 2023
Vanmassenhove E (2020) On the Integration of Linguistic Features into Statistical and Neural Machine Translation. arXiv preprint arXiv:2003.14324 https://doi.org/10.48550/arXiv.2003.14324
Vincent J (2020) Service that uses AI to identify gender based on names looks incredibly biased / Meghan Smith is a woman, but Dr. Meghan Smith is a man, says Genderify. The Verge. https://www.theverge.com/2020/7/29/21346310/ai-service-gender-verification-identification-genderify. Accessed 17 Jan 2023
Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V (2019) Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5310–5319)
West DM, Allen JR (2018) How artificial intelligence is transforming the world. Brookings. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/. Accessed 17 Jan 2023
Zou J (2016) Removing gender bias from algorithms. The Conversation. https://theconversation.com/removing-gender-bias-from-algorithms-64721. Accessed 17 Jan 2023