Computers in Human Behavior: Artificial Humans 4 (2025) 100145

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans

Gender biases within Artificial Intelligence and ChatGPT: Evidence, Sources of Biases and Solutions

Jerlyn Q.H. Ho [a,*], Andree Hartanto [a,**], Andrew Koh [b], Nadyanna M. Majeed [c]

[a] School of Social Sciences, Singapore Management University, Singapore
[b] College of Integrative Studies, Singapore Management University, Singapore
[c] Department of Psychology, National University of Singapore, Singapore

ARTICLE INFO

Keywords:
Artificial intelligence
Chatbots
Gender bias
ChatGPT
Generative AI

ABSTRACT

The growing adoption of Artificial Intelligence (AI) in various sectors has introduced significant benefits, but also raised concerns over biases, particularly in relation to gender. Despite AI's potential to enhance sectors like healthcare, education, and business, it often mirrors reality and its societal prejudices and can manifest itself through unequal treatment in hiring decisions, academic recommendations, or healthcare diagnostics, systematically disadvantaging women. This paper explores how AI systems and chatbots, notably ChatGPT, can perpetuate gender biases due to inherent flaws in training data, algorithms, and user feedback loops. This problem stems from several sources, including biased training datasets, algorithmic design choices, and human biases. To mitigate these issues, various interventions are discussed, including improving data quality, diversifying datasets and annotator pools, integrating fairness-centric algorithmic approaches, and establishing robust policy frameworks at corporate, national, and international levels. Ultimately, addressing AI bias requires a multi-faceted approach involving researchers, developers, and policymakers to ensure AI systems operate fairly and equitably.

1. Introduction

The use of Artificial Intelligence (AI) has influenced almost every field and industry, from healthcare, to education, to private businesses (De Angelis et al., 2023; George et al., 2023; Mhlanga, 2023), altering and improving the ways in which individuals and organizations live and interact. This is largely due to the exponential growth in the capacity and capability of AI (Schwab, 2024). At the center of this change lies ChatGPT, OpenAI's viral AI language model (Walsh, 2022). AI, with its many variations and forms, has had great positive impact, from ChatGPT's ability to improve clinical services through better workflow and medical decision-making (Rao et al., 2023) to enhancing the education experience through personalized learning (Baidoo-Anu & Ansah, 2023; Limna et al., 2023). Given its rapid development and seemingly limitless application possibilities and benefits, the attraction the system has garnered is well-earned. However, much is yet to be fully understood and appreciated.

Despite its revolutionization and the good that it has done, technology should always be treated objectively and periodically reviewed. Unfortunately, both the data that is used to train AI systems (Biswas, 2023) and the codes written by human engineers are sources of bias, as engineers have their inherent bias (Gutierrez, 2021) that is likely to pass onto algorithms if left unchecked. Already, this is happening with other AI systems operating in the healthcare, education, and criminal justice system (Chinta et al., 2024; West et al., 2019). For instance, predictive policing and "dirty data'' in the United States of America due to the derivation and biased analysis of crime data by AI (Richardson et al., 2019) has resulted in a cyclical trend of over-policing of often marginalized communities (Alikhademi et al., 2022). An American Civil Liberties Union report found that significantly more Black residents in Chicago were unlawfully stopped by police as compared to non-Black residents (West et al., 2019). When these marginalized communities are subject to more stops, predictive AI uses such data to forecast future criminal activity, even when police stops do not necessarily reflect actual crime rates. Predictive policing models inform law enforcement on where to allocate resources, which often results in increased police presence in the same areas with more marginalized communities, leading to high likelihood of arrests and crime reports. This creates a

* Corresponding author. Singapore Management University, School of Social Sciences 10 Canning Rise, Level 5, 179873, Singapore.
** Corresponding author. Singapore Management University School of Social Sciences 10 Canning Rise, Level 5, 179873, Singapore.
E-mail addresses: jerlyn.ho.2024@phdps.smu.edu.sg (J.Q.H. Ho), andreeh@smu.edu.sg (A. Hartanto).

https://doi.org/10.1016/j.chbah.2025.100145
Received 3 October 2024; Received in revised form 13 March 2025; Accepted 19 March 2025
Available online 24 March 2025

feedback loop where the algorithm's predictions are continually validated by the influx of data generated from these over-policed areas. This cycle results in even more unfair predictive policing practices that continue to disproportionately impact marginalized communities.

Concurrently, AI systems have been reported to automatically compute better grades for individuals who fit the profile of historically well-performing students (National Audit Office, 2019). That is to say that AI systems might not adequately assess individual performance based on their actual work, but instead have been shown to rely on patterns that align with historical trends (Ovchinnikov, 2020). For instance, they have been shown to favor individuals that come from certain demographics or backgrounds that have historically received preferential treatment, with more resources to perform better (Baker and Hawn, 2022). By relying on historical trends rather than individual merit, AI systems may undervalue students who do not fit the traditional high-performing profile, even if their actual performance is strong.

1.1. Gender bias in AI models

Among the myriads of harmful biases that AI perpetuates, gender bias is one of the most insidious. Unfairness with respect to gender in the face of AI is commonly regarded as a "systemic, unfair difference in the way men and women are treated" (Masiero & Aaltonen, 2020). This insidious prejudice perpetuated by AI in many fields has resulted in women being penalized simply because of their gender (Masiero & Aaltonen, 2020). It can influence hiring decisions in subtle ways, limiting opportunities for women or reinforcing stereotypes during the hiring process (O'Connor & Liu, 2023). Amazon's attempt at developing an automated hiring algorithm was found to inherit sexist sentiments from its training data, resulting in the company launching an internal audit that led to the eventual abandonment of the project (Dastin, 2018). This is largely because the model was trained to screen applicants by observing data created from accepted resumes and highly rated applicants, which came mostly from applicants who were men rather than women (Yarger et al., 2020). From its training, the algorithm learned to favor candidates who were men over candidates who were women (Dastin, 2018; Kodiyan, 2019). This biased decision making further reflects historical gender imbalances in hiring decisions. Hiring bias in AI can extend beyond Amazon and may be especially evident if hiring algorithms are used in and trained on data from similarly historically male-dominated industries.

This bias was not just a reflection of Amazon's own male-dominated workforce (Amazon), but also mirrors the broader industry, as the majority of AI developers are male (West et al., 2019). In particular, in 2018, only 10–15 % of AI developers and designers at major tech firms were female (Whittaker et al., 2018, pp. 1–62). In 2019, only 26 % of the data and AI positions in the workforce as a whole were made up of women (World Economic Forum, 2019). This gender proportion has remained relatively consistent, with 93 % of professional developers in a 2022 Stack Overflow Developer Survey identifying as male (Stack Overflow, 2022). The disproportionate percentage of male developers is a likely cause for analytical and decision-making differences in AI models, since AI systems developed by male engineers may reflect gender differences present in decision making processes. For instance, studies have shown that females tend to engage in more complete information gathering before proceeding on a problem, as compared to males, who tend to proceed with the first viable route (Meyers-Levy & Loken, 2015). Researchers have also shown that females tend to be more risk-averse than males (Burnett et al., 2016), and process decisions to trust differently from males (Riedl et al., 2011). These sex differences may be mirrored by AI decision-making models created by male developers, causing them to emulate male decision-making patterns. As a result, AI systems may reinforce gender biases and fail to account for different perspectives in decision-making processes. Such biases have been exhibited by programs used for hiring, where similarly to male hiring managers (Bosak & Sczesny, 2011; Rice & Barth, 2016), AI models have been reported to favor male candidates, and are reflective of gender imbalances in the tech industry and its resulting AI models (Nuseir et al., 2021).

Apart from the skewed AI developer gender demographic, this hiring discrepancy is also possibly due to the historical data the AI models were trained on. That is to say that if the dataset predominantly consisted of male applicants who had been hired in the past, the algorithm would learn to favor similar profiles, resulting in the replication and reinforcement of existing gender disparities (Bolukbasi et al., 2016;West et al., 2019). For instance, hiring algorithms tend to choose the male candidate over his female competitor, even if both parties have the same qualifications (Langenkamp et al., 2020). The increasing use of AI in recruitment in private corporations (Karaboga & Vardarlier, 2021) has shown that companies that rely on algorithms to hire are likely to be prone to discriminatory hiring along gender lines (Chen, 2023). AI-powered gender biases can impact society through unfair hiring algorithms that systematically disadvantage women by limiting access to job opportunities and career advancement.

The influence of biased training data extends beyond hiring algorithms and into generative AI models, where it can subtly reinforce gender stereotypes through user interactions. For instance, studies have shown that ChatGPT's responses regarding what constitutes as 'good' or 'bad' leadership exceedingly exemplified male leaders as role-models, whilst emphasizing stereotypically masculine traits like courage and risk-taking as ideal characteristics (Newstead et al., 2023). Such responses are modeled after data generated by human responses, where traditionally masculine traits are exemplified, and are themselves a product of gendered and biased labeling (Due Billing & Alvesson, 2000). This pattern is a direct result of the chatbot learning from human-generated content, where masculine traits have historically been associated with leadership. The issue becomes particularly apparent in language-based tasks, where ChatGPT has demonstrated gender biases even in simple, unambiguous queries (Ortega-Martín et al., 2023). In tests done on the chatbot, where it was given a task to detect ambiguity in a sentence that assigned traditionally male occupations to a female pronoun and vice versa, the chatbot failed to put grammar over gender bias, falsely detecting the sentence to be unclear (Ortega-Martín et al., 2023). This occurs because GPT-3 considers the sentence to be confusing in its word choice or sentence structure, suggesting that it is hard to interpret, even if grammatically correct. For instance, when tested with sentences like "the physicist hired the secretary because she was overwhelmed", the chatbot incorrectly assigned the female pronoun to the secretary, even though the sentence structure indicates otherwise. When the gender pronouns were swapped, this error did not occur. Such mistakes suggest that the model has internalized stereotypical gender roles, associating certain professions with men by default.

In education, gender biases in AI systems can lead to unequal treatment and outcomes, particularly for female students. Since dropout prediction by AI may be used to guide educational institutions' decisions regarding acceptance into university courses, any gender biases in such predictions are disadvantageous as they may unfairly reduce opportunities for female students. For instance, Gardner et al. (2019) found that AI algorithms used to predict online course dropout rates performed worse for female students than for their male counterparts. Their study found that the AI model's discriminatory behavior was largely due to the imbalanced gender demographic within courses that it was trained on, resulting in bias towards dominant subgroups that are present in the training data. For example, models trained for science, technology, engineering, or math courses (STEM) courses data tended to be biased to males, since these courses have a higher proportion of male students, as compared to females (Bowman et al., 2022). This discrepancy suggests that the models were not equally effective across genders, leading to inaccurate predictions and subsequent disadvantageous consequences towards female students. Beyond dropout predictions, AI models have also been found to rank intelligence levels along gendered lines (Singh & Ramakrishnan, 2023). For instance, when tasked by researchers to

generate intelligence rankings, the chatbot consistently ranked males of each race above their female counterparts. This is despite consensus and evidence that there are no gender differences in intelligence (Halpern & Wai, 2020). Some instances have shown that the chatbot's answers play into typical gender stereotypes, labelling men as being more logical and less creative (Singh & Ramakrishnan, 2023).

Dropout rates reflect negatively on a university's reputation, which can lead to other downstream consequences regarding funding, finances, and future applicants (Marcinkowski et al., 2020, pp. 122–130). To minimize such risks, universities seek to minimize such risks by making prudent decisions regarding their admissions. Since studies have shown that AI systems disproportionately predict higher dropout risks for female students, especially from STEM programs, based on biased data (Gardner et al., 2019), these predictions could then influence universities to accept fewer female applicants to certain programs, such as STEM ones. Such scenarios perpetuate gender disparities in higher education access, particularly in traditionally male-dominated fields. This can lead to a cycle of exclusion and continuation of gender gaps in male-dominated industries. Academic recommendations done by AI models have also been proven to tend to assume that male students are more capable than female students by consistently advising female students to enroll in foundational courses, whilst recommending more advanced ones for their male counterparts. This is especially so for STEM courses (Khan, 2023). The biases observed could lead to systematic disadvantages, further entrenching the educational gender gap.

In healthcare, while the incorporation of AI has been shown to broadly improve the accuracy of diagnosis and treatments (Karalis, 2024; Kumar et al., 2023), it has also introduced gender biases that disproportionately affect women's health outcomes. These biases stem from AI databases and forecast models trained primarily on male-dominated datasets (Hamberg, 2008; Perez, 2020), leading to systemic discriminatory healthcare services and treatment inequalities along sex lines (Cirillo et al., 2020). For instance, female patients' chronic heart disease symptoms were shown to be significantly more likely to be misdiagnosed and misattributed to other conditions (e.g., gastrointestinal conditions) than males, even when the same symptoms were reported (Maserejian et al., 2009). Furthermore, researchers found that diagnostic AI meant to specialize in diagnosing different thoracic diseases in women was tested to in fact be better at doing so for men (Ganz et al., 2021). Diagnostic programs were also completely inaccurate, resulting in disproportionately more underdiagnosis, or even misdiagnosis for women (Seyyed-Kalantari et al., 2021). If an AI model designed to address women's health issues performs better on male patients, it suggests that the data used to train it is biased or unrepresentative of the female population. Since women have been historically more likely to be under or misdiagnosed, recommended treatment plans for women will likely follow this similar trend (Cirillo et al., 2020; Hamberg, 2008). AI systems trained on this biased data are highly likely to replicate and reproduce these disparities. Such biases have already been proven to occur in treatment plans recommended by GPT-4, where studies done to measure the sex difference in likelihood of the chatbot recommending treatments and further clinical testing showed that it was significantly more likely to do so for males than females (Zack et al., 2024). Consequently, AI in healthcare can lead to women continually experiencing misdiagnoses, or incorrect treatment pathways.

2. Chatbot design

This track record of AI being mishandled and unrepresentative is worrying given its increasing pervasiveness. As noted, a particular area of concern is the rise of large language models and AI chatbots, notably ChatGPT. AI large language models such as ChatGPT are built on an intricate architecture of the most advanced AI capabilities including Natural Language Processing (NLP), Machine Learning (ML) and deep learning (Haleem et al., 2022). Such chatbots' intersection of human linguistics and AI is possible due to its intensive learning and training from massive amounts of corpus data (Feng et al., 2023), a process otherwise known as ML (Ben Letaifa, 2019). By using algorithms to learn from massive amounts of data (Hassani & Silva, 2023), chatbots such as ChatGPT equips itself with the ability to learn and mimic human language (Cai et al., 2023). Specifically, ChatGPT utilizes a deep neural network based on the Transformer architecture (Vaswani et al., 2017), which allows for greater processing efficiency and accuracy in understanding language structure (Radford et al., 2018). At the core of its design is the two-stage learning process involving both supervised[1] and unsupervised[2] learning, which allows the chatbot to recognize and identify patterns between data samples, before extrapolating this new knowledge onto new dataset to further its learning (Alzubi et al., 2018). Supervised learning models are trained with specific input and output labels that allow for more accurate prediction as per its training data (Jiang et al., 2020). For ChatGPT, such training happens alongside unsupervised pre-training, where the model is fed large amounts of data to read and learn how to predict the next word in a sentence (Radford et al., 2018). By leveraging on both types of learning, ChatGPT is able to achieve even higher accuracy in its prediction, whilst being more proficient in discovering patterns in data (LeCun et al., 2015). This combination is a critical part of the technology behind its ability to generate accurate responses to queries (Nasteski, 2017).

As the model completes the learning phase, it also undergoes fine-tuning, which broadly improves the model's overall performance and generalization, especially in conversational settings (Radford et al., 2018). This generally involves training the model on labeled datasets where both inputs and outputs are provided, and enhances the model's ability to generate coherent, relevant, and task-specific responses (Radford et al., 2018). Apart from this, ChatGPT's advanced NLP model allows it to comprehend and generate human-like, natural-sounding text (Chowdhary, 2020) when responding to questions. The resulting combination of this rigorous design and learning pipeline is an AI chatbot capable of human-like interactions (Brown et al., 2020) whilst delivering mostly accurate answers pulled from its database.

3. Chatbot biases

Just a year after its launch, the use of ChatGPT in our everyday lives has been widely accepted (Huallpa et al., 2023; Limna et al., 2023; Temsah et al., 2023). With its ability to provide information on vaccinations, disease detection, and cancer-related queries (Biswas, 2023; Li et al., 2023), this high penetration rate (Hu, 2023) signals the high level of trust that the general populace has in the chatbot (Choudhury & Shamszare, 2023; Sun et al., 2024). However, ChatGPT's proliferation and influence has rightly sparked concern regarding the ethics behind its use (Dwivedi et al., 2023), as well as worries surrounding its objectivity (Wu et al., 2023). Despite concerns surrounding ChatGPT and its potential to reproduce and perpetuate biases, research targeted at gender-related pitfalls has only emerged even more recently. Furthermore, much of the existing literature on AI bias has focused on predictive models built from supervised machine learning algorithms. However, research on bias in generative AI models remain underexplored, particularly those used in chatbot applications. The present narrative review seeks to extend the literature beyond the realm of bias in supervised machine learning models to highlight sources of bias in the design and implementation pipeline of generative tools such as ChatGPT. This review will contribute a synthesis and integration of interdisciplinary sources, including peer-reviewed articles, industry reports, and policy documents. By doing so, we aim to provide a broad,

[1] Supervised learning uses a training set to teach models to yield a desired output. This training dataset includes inputs and correct outputs that have been labeled, which allow the model to learn over time (Google Cloud, n.d).

[2] Unsupervised learning uses unlabeled data from training sets. From that data, the model discovers patterns and associations (Google Cloud, n.d).

multifaceted perspective on the issue at hand. Hence, the current review will be conceptual and narrative in nature, to discuss existing literature on gender bias in AI systems, with a focus on theoretical discussions and emerging empirical evidence, rather than a systematic one. We believe that by examining biases present in training data, model architecture, and user interactions, this review will provide a comprehensive framework for understanding and mitigating gender bias in AI systems.

Indeed, despite AI's significant progress and increasing intertwinement with our everyday lives, AI systems are not perfect. Part of the initial goals of AI systems were for them be an embodiment of objectivity (Alkurd et al., 2020; Restrepo Amariles & Baquero, 2023; Waseem et al., 2021) and thus the solution to innate human biases (Batra et al., 2022; Houser, 2019; Yudkowsky, 2001). Whilst society embraces AI and its potential, it is important to do so with an understanding that interactions with such systems are a reflection of their human engineers, whose biases may be reflected during their development of these systems (Martínez et al., 2022; Marinucci et al., 2023). Further, since AI models draw their knowledge and training from massive amounts of data and code (Kocoń et al., 2023), AI chatbots such as ChatGPT can therefore provide answers that are reflective of any biases and representations that exist in the dataset (Biswas, 2023). Additionally, the majority of AI systems that chatbots are developed from consists of data and algorithms that have interacted with human developers (Bender et al., 2021; Caldarini et al., 2022; Xiang, 2024). As a result, subjecting these systems to any biases or discriminatory sentiments exhibited by these developers, and may even be amplified by AI and chatbots (Mohanani et al., 2018; Xiang, 2024).

Broadly, there are several ways in which these biases can infiltrate AI systems (Fig. 1), impacting the fairness and accuracy of its responses. Fundamentally, AI chatbots rely on corpus data to be trained. If the data at this stage is already biased, existing discriminatory sentiments will worm their way into systems that rely on it, creating biased algorithms (Nadeem et al., 2020; Obermeyer & Mullainathan, 2019). Specific to chatbots, if they are trained primarily on datasets that reflect societal biases, there is a high likelihood of internalization and downstream reproduction of such stereotypes in their interactions with users (Bender & Friedman, 2018). Consequently, the chatbot's responses will not only reflect existing biases, but are likely to perpetuate such sentiments by affecting users' interactions and responses (Vicente & Matute, 2023). Moreover, when users react apathetically or even positively to such biased responses, such interactions reinforce inherent biased learnings within the algorithm (Navidi & Landry, 2021), forming a new basis of learning that will inevitably be taught to future systems (Navidi & Landry, 2021; Wu et al., 2022). This cyclical loop of learning, interaction, feedback, and dataset creation is where the root source of where biases in chatbot begin, and where it must be disentangled from. Given that ChatGPT is presently the most widely used and accessible chatbot (Huallpa et al., 2023; Limna et al., 2023), the following sections and subsections will employ the chatbot as a running example to better illuminate our key points, where applicable. The following sections details the various biases present in each of the following processes: training data, chatbot architecture, and user feedback loop, as well as how these biases are manifested.

3.1. *Biases within training data*

Most biases begin at conception. In the case of AI and Large Language Models (LLM), this refers to the training data and datasets that the chatbots are provided with (Feine et al., 2020; Bradley and Alhajjar, n. d.). ML models are limited by the quality of their associated training data (Daneshjou et al., 2021) and are subject to bias or inaccuracies present in its dataset. Many stereotypes, including gender and racial ones, are often sown at this stage (Tejani et al., 2024). Oftentimes it is the quality, diversity, and representativeness of training data that can shape biases and downstream lack of objectivity and accuracy in chatbots' responses (Norori et al., 2021). Biases in data can manifest and be traced back to various sources. Although training data can consist of other pitfalls such as user inputs (Fig. 1), the following details only the most prominent types of biases present in datasets that LLMs train on.

3.1.1. *Representation bias*

Starting with **representation bias**, this arises when the AI model is presented with, and learns from, incomplete and non-representative datasets (Norori et al., 2021) which often lack the nuances, diversity,

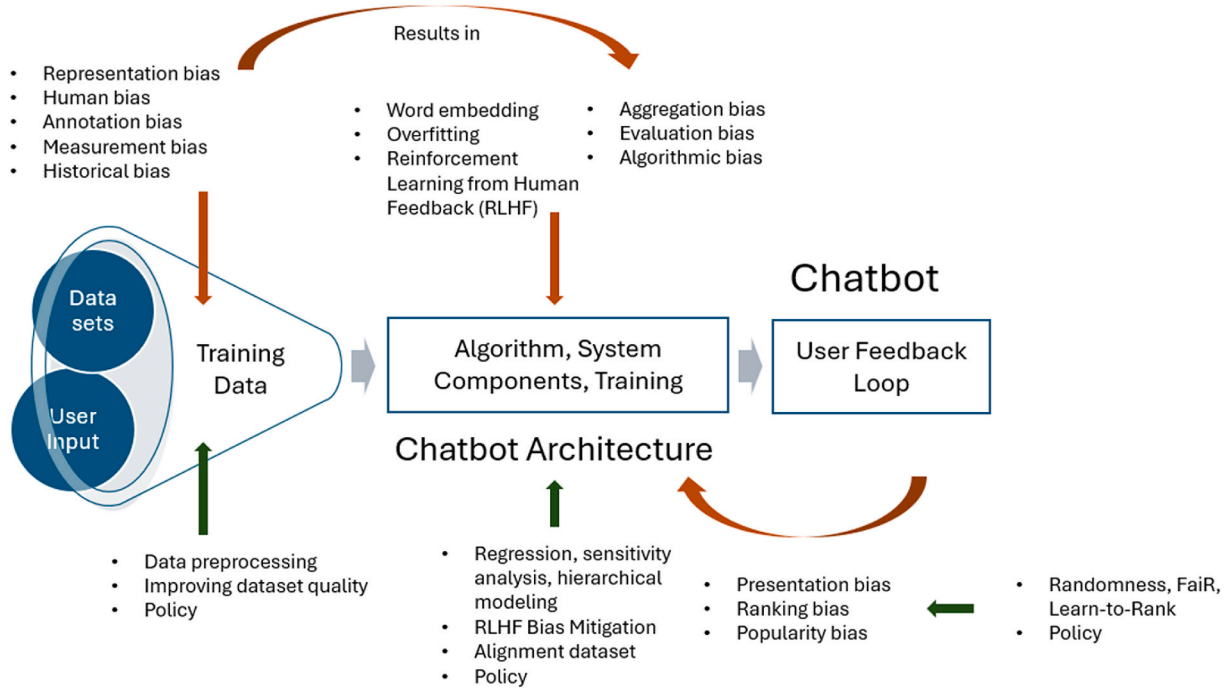

Fig. 1. Types of Biases in Chatbots
Note. This figure demonstrates where each type of bias is present in chatbots as well as their accompanying mitigation strategies.

and subtleties of the actual populace (Norori et al., 2021; Shahbazi et al., 2022). Furthermore, since datasets are more likely to be centered on or otherwise produced by the dominant culture, gender, or language, this means that marginalized demographics are likely to be under- or mis-represented (Kaiksow & Carter, 2022; Woodall et al., 2010). Downstream consequences include negatively impacting the machine's ability to generalize well for under-represented demographics. This lack of representativeness emerges when datasets are outdated (Bandy, 2021; Zhuo et al., 2023), overly generalized (Lloyd, 2018), or assume heterogeneity in the populace (Obradovic & Vucetic, 2004, pp. 1–24). Assumed data heterogeneity occurs when the data is thought to be diverse and representative, even if this is not the case. When this occurs, certain demographics may be underrepresented, resulting in a biased dataset. In cases of representation bias, the chatbot's algorithm is likely to have a higher tendency to pick up and assign learned generic characteristics to all associated with the group (De-Arteaga et al., 2019, pp. 120–128), whether accurate or otherwise. For instance, financial chatbots may offer different solutions based on stereotypes stemming from datasets; they may prioritize or offer higher-cost products to users based on their assumed financial status that is congruent with gender or racial stereotypes present in the non-representative dataset it was trained on (Riyazahmed, 2023). In the case of ChatGPT, the training data includes a disproportionate amount of text reflecting male-centric perspectives. As a result, when asked about leadership, ChatGPT often associates traditionally masculine traits such as courage and assertive, with effective leadership, while neglecting to highlight nurturing or collaborative traits, which are stereotypically feminine traits (Newstead et al., 2023).

3.1.2. Human biases

Such non-representative training sets first arise from biased datasets that stem from and were affected by **human biases** during its inception of reporting and collation (Gordon & Van Durme, 2013, pp. 25–30; Winship & Mare, 1992). Human involvement in data collection introduces human biases such as sampling, in-group, and measurement biases, which affect how the data is handled during the selection and collation process (Ruiz-Gutierrez et al., 2016; Pannucci & Wilkins, 2010). These biased datasets are then turned into datasets and fed as AI training data. These human biases often go unnoticed, making their way as flawed training data used by AI systems, further compromising its objectivity.

3.1.3. Annotation bias

One of the most prevalent human biases in the context of AI is otherwise known as **annotation bias** (Al Kuwatly et al., 2020). Data annotation refers to the process of labeling data to make it comprehensible to the LLM, with the annotation process being handled by humans (Ji et al., 2024). Biases during this process happen when the personal prejudices, subjective stereotypes, and discriminatory tendencies of the human annotator influences the data's label (Beck et al., 2020). This process introduces inherent human biases into data, which is then absorbed into algorithms and reproduced during user interactions (Bender & Friedman, 2018). Chatbots have been observed to emulate this biased language and responses that it was exposed to during this annotation process (Nasution & Onan, 2024). Studies have shown that user interactions with chatbots influenced by annotation biases are often colored by commonly held biased sentiments (Dinan et al., 2020; Wan et al., 2023). For instance, when asked which sex was smarter, Microsoft's commercial chatbot was confident in its answer of "men" (Wan et al., 2023). This occurs partly because the group of annotators working on a dataset is not diverse enough, causing their biases to be captured by AI models and thus leading to issues such as those regarding gender bias (Geva et al., 2019).

3.1.4. Measurement bias

Relatedly, **measurement bias** arises when the choosing and utilization of data labels for predictions are inappropriate (Schwartz et al., 2021). This means that the label on data points may be used as a proxy to predict adjacent variables that these data points do not accurately represent (Varona & Suárez, 2022), which may result in inaccurate estimations (Zinn et al., 2021). Predictive policing in America is one of the most egregious consequences of measurement bias. This happens when AI systems inaccurately use arrest rates—rather than true crime—as an indicator of safety (Richardson et al., 2019). In this context, although adjacent, arrest rates can be a consequence of racial bias, since arrest rates for non-white individuals have been shown to be disproportionately higher (Barnes et al., 2015). By using arrest rates as a proxy for safety, AI systems may wrongly deem areas with non-white majority communities as more unsafe as compared to their white majority counterparts.

3.1.5. Historical bias

Even if data is collated and collected perfectly, another problem pertaining to flawed datasets is **historical bias**. This happens when the state of the world when the dataset was generated is no longer comparable to the present (Shahbazi et al., 2023). That is to say that the state of the world and its societies have profound impacts on the data that it generates. For instance, data taken from twenty years ago pertaining to the number of women in high-ranking positions in both private and public offices will no doubt be significantly smaller than that of present day's data (Berry & Franks, 2010; Goryunova & Madsen, 2024). Although the data generated may be historically accurate, biases and consequences may arise when this data is being used to predict and inform in the present (Naci & Soumerai, 2016). This becomes an insidious problem should the AI model take this historical data as true of the current state of the world.

3.1.6. Implications of dataset biases

Given that datasets used to train chatbots have been proven to be non-objective, it is no surprise that biases present in training data often resurface during user interactions. In particular, ChatGPT has been proven to succumb to such biases (Brown et al., 2020). Training data fed into ML models such as GPT-3 typically come from publicly available data such as the internet, articles, and books (Johri, 2023), all of which originate from human creators. In the realm of politics, media coverage has been shown to mention female politicians more than males, seemingly giving female politicians more attention than their male counterparts (Leavy, 2020). However, on closer inspection, these articles typically contain stereotypical language and associations, whilst portraying a happier family life for male politicians compared to females (Leavy, 2020).

Such biased coverage and language often find its way into downstream responses when interacting with users. For instance, when faced with a prompt to identify traditionally "highly educated" occupations such as professors, GPT-3 consistently associated the role with male identifiers, rather than female ones (Brown et al., 2020). Kaplan et al. (2024) also concluded that the chatbot reproduced gender-based language in its writing of recommendation letters, often using stereotypical language when tasked to write letters for female names. In particular, their study showed that letters generated for female names consisted of more social and communal language such as "friend" and "family", while letters for males emphasized their professional skills (Kaplan et al., 2024). The authors suggested that this was due to the training data that ChatGPT was subject to, wherein the English language corpora widely categorizes females in the same aforementioned social-affiliated language and tone (Kaplan et al., 2024), reproducing similar stereotypes when prompted.

Dataset biases can significantly influence the behavior and responses of chatbots. When training data reflects societal stereotypes or underrepresents certain demographics, chatbots will learn to replicate as well. Apart from biases present in the training data, it is also important to note that the architecture of chatbots also plays a crucial role in further entrenching such biases. The next section explores the underlying

structure of chatbots, and how this architecture underpins other forms of biases present in AI and chatbots.

3.2. Biases within chatbot architecture

Although chatbot architecture typically refers to the technical and structural design of the model, system components like the developers behind these processes have been categorized under this section (Fig. 1). LLMs such as ChatGPT (Arcila, 2023) consist of transformer-based neural network architecture that is used for NLP (Aydin & Karaarslan, 2023). Otherwise known as the training stage, the many layers of the chatbot's neutral network (Wu et al., 2023) helps the model "understand" human language by learning patterns and relationship between words (Aydin & Karaarslan, 2023), that were gleaned from its training data. During this phase, the algorithm learns to make accurate predictions (Wu et al., 2023) as well as understand the relationship between words (Wu et al., 2023), forming the basis of its responses during interactions with human users. In the case of AI, it is at this training phase that chatbot responses are likely to learn to mimic and recreate stereotypes (Ntoutsi et al., 2020). Broadly, its infrastructure can further entrench biases, turning it into a systemic problem that becomes increasingly difficult to disentangle from the model. The following details how such biases get embedded in the system in the first place, and how they may persist and manifest in user interactions.

3.2.1. Word embedding

A critical part of this process that introduces non-objectiveness, whilst having a great impact on downstream consequences is **word embedding**. This is a key component of NLP models (Mikolov et al., 2013) such as ChatGPT, and works by capturing semantic relationships between words, allowing the model to form connections between similar or associated words (IBM, 2024). However, because algorithms and AI models do so with numbers rather than actual contextual information or knowledge (Worth, 2023), the model can be prone to biases and stereotyping. Through the labeling process, algorithms have been shown to assign static, non-contextual properties along racial and gender lines (Bolukbasi et al., 2016; Marinucci et al., 2023). Such subjective word embedding has been observed to favor majority population groups, namely that of gender, race, and language (Zhang et al., 2020, pp. 110–120). For instance, researchers have shown that ML systems tend to associate science and art terms according to gender stereotypes (Angwin et al., 2016; Caliskan et al., 2017). Due to such stereotyping, it is generally regarded as inaccurate, with great implications on downstream tasks (Borah et al., 2024) such as interactions with chatbot users.

3.2.2. Aggregation bias

At this development stage, **aggregation bias** may also arise, which occurs when different data groups or distinct populations are inappropriately combined by some arbitrary variable (Vázquez-Ingelmo et al., 2020). This then results in a model that only performs well for a majority group, whilst giving suboptimal performance when presented with non-majority groups (Landau, 1978). That means that the generated algorithm is a one-size-fits-all, blanket model that does not consider the heterogeneity of a populace (Vázquez-Ingelmo et al., 2020). For instance, if crime rates were aggregated based on city-level geography, one may mistakenly assume that because a certain city is deemed safer compared to another, all parts of the safer city must be so. The overall assumption that a city deemed "safer" is uniformly safe ignores the fact that there are neighborhoods within the "safer" city that may be more dangerous than neighborhoods in a city considered "unsafe". The oversimplification of the data results in a misleading conclusion because distinctions between neighborhoods are overlooked when the data is aggregated at a higher, city-wide, level. Effectively, this is similar to representation bias, only translated into model and algorithmic forms.

3.2.3. Evaluation bias

Evaluation bias arises during the development stage as the model is often optimized using specific key metrics and characteristics that may not be generalized to the rest of the general populace (Fahse et al., 2021). That is to say that during evaluation and optimization of the model, its quality and accuracy are often judged and measured against certain benchmarks (Fahse et al., 2021). However, biases develop when these benchmarks do not accurately represent the populace as a whole (Fahse et al., 2021). For instance, facial analysis benchmark datasets were shown to be composed primarily of lighter-skin subjects, and that performance of the model has been tested to be highly accurate when tested on subjects with similar benchmarks. However, the same model underperformed when tasked to do the same with non-white subjects (Buolamwini & Gebru, 2018). This means that ML models may perform accurately only when used and tested on the same benchmark, but falter when tasked to do so with another demographic (Chu et al., 2024; Neville et al., 2012). Oftentimes, models are inaccurately evaluated as objective, especially if the process of testing against other benchmarks is overlooked during the training phase.

3.2.4. Algorithmic bias

Like all ML models, **algorithmic bias** can arise during its development phase. This occurs when the algorithm takes the liberty to make biased designs and decisions based on what function it was meant to perform (Fahse et al., 2021). That is to say that the algorithm itself is biased, with its root cause being a biased dataset (Akter et al., 2022) from which it learns from. However, that is not to say that a perfectly unbiased dataset will yield a non-biased algorithm. For instance, social media algorithms are optimized to keep users on its platforms by keeping them engaged (Manoharan, 2024). However, rather than showing users topics that they may be interested in, algorithms have learned that user engagement can also be driven by hate and antagonistic content (Rathje et al., 2021). In this case, while the algorithm has been optimized to perform its function well, it does so by playing into personal prejudices and discrimination, further perpetuating and reproducing biases. Mauro and Schellmann (2023showed how insidious algorithmic biases can be in terms of sexist and gendered contents, where AI algorithms were proven to suppress and over-sexualize content about women. In particular, they highlighted how AI classifiers wrongfully analyzed a bra to be inherently "racy" when tested on a male participant who was previously half-naked, and was not analyzed to be "racy" without the bra on (Mauro & Schellmann, 2023). Consequently, algorithms may deem similar images—of women simply wearing a bra—as sexual and suggestive, whereas their half-naked male counterparts are not.

3.2.5. Overfitting

LLMs are also susceptible to **overfitting** (Jebali et al., 2024), a phenomenon which occurs when an AI model learns not only the underlying patterns in the training data, but also takes into account noise and outliers (Ying, 2019). This is typically caused by models being overly complex, such as having the presence of too many or complicated parameters, or simply because the dataset it has been trained on is insufficient or unrepresentative (Ying, 2019). In other words, an LLM that has been overfit may be seen as memorizing and regurgitating specific phrases rather than learning the underlying patterns behind its training data. This technical, algorithmic issue causes the model to perform well only on the data it has been trained on, but fails to generalize as well when faced with new or unseen data (Mutasa et al., 2020). Overfitting in LLMs can be harmful, as it may be overfit to biases present in datasets used for its training. That is to say that it may memorize these biased patterns and reproduce similar sentiments when interacting with users.

3.2.6. Reinforcement Learning from human feedback

LLMs, of such large scales such as ChatGPT, are often subject to a 2-

stage training process comprising of unsupervised pre-training and **Supervised Fine-Tuning** (SFT) (Ji et al., 2024; Roumeliotis & Tselikas, 2023). During these stages, the model first learns the underlying patterns, before it is tweaked to further improve its performance (Lund & Wang, 2023). A third, final step during this development stage is known as **Reinforcement Learning from Human Feedback** (RLHF). SFT leverages labeled or annotated data to passively optimize and finetune the LLM, which will then be refined and adjusted accordingly (Ji et al., 2024). This is repeated and reiterated as many times as needed until a specific accuracy metric is met for whatever function it is meant to achieve (Ji et al., 2024). The SFT process can be heavily impacted by annotation biases within the dataset. Similar to SFT, the intent of RLHF's process aims to fine tune LLMs and optimize it for user interactions (Li et al., 2024). Dissimilar to SFT, the RLHF process entails the presentation of its tasks and outputs to humans for evaluation, feedback and eventual improvement of performance (Liu, 2023). This feedback can come in the form of reward models or annotations (Ziegler et al., 2020), which opens the model up to innate human biases and personal prejudices from its human developers involved in this step. For instance, when evaluating AI models' performance in terms of alignment with human preferences, researchers found that models trained with RLHF produced less diverse responses, instead generating outputs that were more closely aligned with human feedback. This suggests that biases introduced in the RLHF process are reflective of human inputs and innate preferences (Kirk et al., 2023).

3.2.7. Implications of biases within the chatbot's architecture

Instances of architecture-based gender biases are most prevalent and distinct in religion-based AI chatbots akin to ChatGPT. Chatbots such as BibleMate.org, GitaGPT, and QuranGPT (Lee, 2023; Nooreyezdan, 2023; Wright, 2024), were launched with the hopes of making their respective religions more accessible, whilst providing guidance to its users (Biana, 2024) and relevant believers. Prominently, the Hindu-based LLM of GitaGPT has been proven to justify acts of violence; in particular, murder and misogyny were condoned by the religious bot (Nooreyezdan, 2023). Unfortunately, such responses were not unique to GitaGPT, and were in fact prevalent in other religious LLMs (Biana, 2024). This is widely because of the algorithm generated by the religious text that was fed to the chatbot, wherein the architecture of the chatbot merely extrapolates religious contents, without any contextual details that may have been present in the texts (Biana, 2024). Broadly, these consequences were a result of complacency at this stage, resulting in sexist, dangerous responses by the chatbots. Although it is difficult to determine if gender-biased outputs by chatbots are purely due to biases at the architecture stage, or a combination of flaws cascaded down from the dataset, flaws at this stage are sure to exacerbate and amplify any existing biases present. Since datasets are largely prone and susceptible to biases, one can assume that ML models and its NNs will intensify and further integrate any gender biases into its system, making it harder to disentangle and decouple.

In summary, the architecture phase of chatbot development involves the design of underlying systems, algorithms, and ways of learning that enable the chatbot to function. This phase is important as it informs the chatbot on how to process language and interact with users. Without careful consideration of potential biases present in the algorithm and system's design, the architecture can unintentionally perpetuate bias in its interactions.

3.3. Biases in the user feedback loop

User generated biases are also a large reason behind biases in the ML models, specifically LLMs such as chatbots. Although not as straightforward as biases originating from datasets, training processes, or the development phase, interactions with users after deployment can also lead to chatbots picking up on innate human prejudices and inclinations (Silberg & Manyika, 2019). A chatbot's inherent purpose is to interact with users—a process that oftentimes entails a back and forth of prompts and responses between both parties, till a satisfied result is returned to the human user. This communication inadvertently relays and recycles information containing preferences, biases, and proclivities (Silberg & Manyika, 2019). The LLM's algorithm will eventually pick up on this behavior and introduce biased responses to its next user. There are several ways in which this vicious cycle is reproduced in the chatbot-user feedback loop.

3.3.1. Presentation bias

In cases of chatbots like ChatGPT, with functions similar to search engines, an integral part of its role includes scouring and presenting information (Haleem et al., 2022) as a response to queries. However, because of the sheer volume of information on the Internet, it is almost impossible for human users to view, and for the algorithm to present every piece of information individually (Heylighen, 2002), herein paving the way for **presentation bias**, where information presented first and seen by users are more likely to be viewed and clicked into (Baeza-Yates, 2018). For instance, Google's extensive search engine returns multiple pages per query, yet most users admit to only viewing information presented on the first page, with only a few clicks to the second (Barry & Lardner, 2011). While the top results are not necessarily the most accurate or reliable, they gain visibility simply by being ranked higher in the results (Baeza-Yates, 2018). This visibility prompts user interactions that lead to further feedback—both positive and negative—that further reinforces the prominence of the content (Altaf, 2019). As a result, the algorithm then continues to prioritize and recycles this content onto other users' searches. Since it has been established that algorithms have been shown to prioritize antagonistic and discriminatory content over accurate ones (Rathje et al., 2021), this feedback loop poses significant risks, as the model may increasingly promote harmful and biased results.

3.3.2. Ranking bias

Ranking bias often occurs in tandem with presentation bias, and stems from the human tendency to favor and assign more importance to the first information encountered. This cognitive bias is otherwise known as primacy effect (Van Erkel & Thijssen, 2016). In the context of ranking bias, users might perceive the first result as the most relevant or accurate, simply because it's the first one they encounter (Bar-Ilan et al., 2009; Lerman & Hogg, 2014). As a result, content ranked first tends to be accessed more than those below (Collins et al., 2018). When this occurs, the same content is inevitably pushed up the rankings, leading to even more clicks (Baeza-Yates, 2018), as the algorithm continues to prioritize what gets the most engagement, perpetuating its high ranking based on its position in the list even if it is not the most accurate. Just like presentation bias, this is an insidious problem as the algorithm may eventually prioritize ranking discriminatory content over objective ones.

3.3.3. Popularity bias

Presentation and ranking bias are closely related to **popularity bias**, which refers to the tendency of algorithms to favor content or items that have already gained significant attention or engagement (Abdollahpouri et al., 2021). The algorithm interprets user interaction with items as a sign of popularity, leading to further exposure and visibility. Since items that are presented more or ranked higher tend to enjoy better traction and click-ins, they tend to be perceived to be "popular" by the model, causing them to remain highly recommended and popular (Zhang et al., 2021), whilst less exposed items stay under-recommended as they fail to generate the same level of engagement (Zhu et al., 2021). This feedback loop leads to a concentration of visibility on a few popular items while other content is sidelined, regardless of quality or relevance.

3.3.4. Implications of negative user feedback - microsoft tay

The most prolific bias stemming from user-interactions is Microsoft's

Tay, one of the earliest renditions of an AI chatbot capable of understanding and having conversation with human users on social media site, *X* (previously known as Twitter) (Vincent, 2016). Meant to mark an evolution in technological capabilities of AI chatbots, Tay was intended to learn from human users on *X* (Mathur et al., 2016). Yet, less than a day later, the chatbot was spewing racist, antisemitic, and sexist comments (Mathur et al., 2016), seemingly learnt from the human users it was meant to interact with. Although its neutral, learning algorithm played a part (Wolf et al., 2017) in allowing the chatbot to learn from all sources, human interactions were the main driving factor in its descent (Wolf et al., 2017). Online users were shown goading and convincing the bot to repeat misogynistic and harmful statements—to which it did, at an alarming fast rate (Neff & Nagy, 2016). Within a day, Tay was shut down by Microsoft (Victor, 2016). Tay's example highlights how quickly AI, and ML models can learn and evolve from their interactions with users. Taken in tandem with other prevalent biases present in the user-feedback loop, chatbots, in particular, ChatGPT, have the potential to turn into a similar phenomenon.

4. Interventions

Since biases arise at every step of the process, and are often re-injected into the model, mitigating these biases requires a much more nuanced and targeted approach at every level. Although solutions are often tailored to each AI model's intended use, there are some strategies that underpin bias mitigation for all models. These multi-pronged solutions require efforts ranging from technological and systemic interventions to policy improvements wherever possible (Fig. 1). The following section details solutions that may be used at each stage of chatbot development.

4.1. Dataset interventions

4.1.1. Data preprocessing

Since a sizable proportion of the issue at this stage involves the lack of diversity in datasets used for training, attempts must be made to remove biases at this stage. **Data preprocessing** helps to ensure clarity regarding the models' intended use and purpose, in turn shaping its preparation process. Broadly, this means the modification of datasets before a model is trained on it (Linardatos et al., 2020). Kamiran and Calders' (2012) approach details a four-pronged method in doing so. Firstly, suppression (Kamiran & Calders, 2012) entails the lowering of impact of specific correlated variables, such as age or race, on the model's decision-making process. This approach compels the algorithm during the development stage, to place less emphasis on such attributes and instead rely more on correlations based on other features. The result is an algorithm that is not biased towards any demographics (Linardatos et al., 2020).

Secondly, the dataset is "massaged", wherein labels of objects are changed or removed with the purpose of removing discriminatory sentiments (Kamiran & Calders, 2009). This deliberate, systemic process is based on the idea of identifying and correcting identified biases present in the dataset in order to create a fairer and more balanced one (Kamiran & Calders, 2009). The process relies on robust statistical analysis and algorithmic assessments to objectively determine where biases exist within the model, and how they affect its performance (Zhou et al., 2021). Once these biases are identified, the labels associated with certain data points are changed in a systematic way to reduce the biased outcome. For instance, negative labels disproportionately present in certain demographics may be changed to attain more balanced representation (Kamiran & Calders, 2012). The algorithm that is trained on this relabeled data does so without any dataset biases present (Kamiran & Calders, 2009). After massaging the data, models should also be validated and tested using fairness metrics to ensure that the changes have reduced bias and improved fairness. Such fairness metrics involve checking against statistical metrics to measure and ensure that the model's performance is consistent across different demographic groups (Lum et al., 2022, pp. 379–389). The debiased algorithm is then rechecked to ensure that these debiasing frameworks are indeed successful, a process that may include running the model through simulations and other statistical coverage checks (Lum et al., 2022, pp. 379–389). The culmination of these steps ensures that decisions made by the algorithm are grounded in empirical data, rather than subjective ones. Then, tuples, which are sets of values representing specific objects in a dataset (Jespersen, 2003), are re-assigned weights (i.e., re-weighted; Calders et al., 2009). By doing so, data can be re-balanced with regards to any attributes, leading to a more independent classification[3] (Calders et al., 2009), without the influence of extraneous variables that might affect the model's decision-making process.

Lastly, over-sampling broadly details increasing the number of datapoints in often underrepresented demographics such as marginalized groups in order to increase the perceived importance of such groups (Kamiran & Calders, 2012). This then forces the algorithm to pay more attention to these groups. This process is also known as the Synthetic Minority Over-sampling Technique (SMOTE) (Kamiran & Calders, 2012). This technique is used specifically to address class imbalance in a dataset, resulting in a more balanced dataset consisting of additional data points from often underrepresented groups. The algorithm is then able to learn patterns and relationships that are representative of all groups, not just the majority one. The preprocessing process of the development of the model aims to mitigate frequent sampling pitfalls such as undersampling, oversampling, or unrepresentative sampling (Qraitem et al., 2023).

4.1.2. Improving dataset quality

As discussed, the root of many dataset biases lies in the quality of its training data. Therefore, a critical intervention is the provision of high-quality data, which can be achieved through the **diversification of data samples** (Neilsen et al., 2017) and **annotator pools** (Prabhakaran et al., 2021). Diversification of data involves sourcing datasets from a broad range of origins rather than relying on a select demographic or database (Neilsen et al., 2017). This diversification process involves ensuring that the dataset and annotators capture a wide range of scenarios, contexts, and variations by including data from different sources. For instance, LLMs should diversify data by including voices from different accents in a speech recognition model or including text from different dialects, languages, and accents. This practice helps to prevent bias and improve the model's ability to handle diverse inputs, ensuring that it can accurately recognize a broader range of accents. By consciously widening the data pool, underrepresented demographics are more likely to be captured in data points, and out-of-date data is more likely to be corrected (Kuhlman et al., 2020). This results in a reduction of representation, measurement and historical biases in the dataset, allowing for better generalization by the AI model.

By diversifying the pool of annotators to include various demographics, perspectives, and backgrounds, individual differences and biases are likely to be mitigated (Prabhakaran et al., 2021). For chatbots such as ChatGPT, this is especially important as datasets and annotator pools should reflect the diversity of its consumers. Not only do these mitigation strategies have a direct impact on the quality of training data, biases at the development stage are also likely to be affected as a positive downstream consequence. For instance, evaluation and word embedding biases are heavily impacted by the quality and diversity of training data. By ensuring and improving upon biases at such an early stage, AI

[3] Independent classification refers to the algorithm's ability to assess and classify data points based on their relevant attributes and features, rather than being swayed by irrelevant factors commonly associated with race, gender, or age. For instance, a hiring algorithm capable of independent classification would qualify candidates based only by their qualifications and experience, without the influence of race, gender, or age.

models are less likely to retain and reproduce such biases in its training phase.

4.2. Architecture interventions

During the training and development phase, an algorithm is generally left to further optimize on its own accord (Ji et al., 2024; Roumeliotis & Tselikas, 2023). However, human developers should use objective and statistically validated debiasing frameworks to monitor and modify algorithms during this process before inherent biases are worked into the system (d'Alessandro et al., 2017; Idowu et al., 2024). Not only can algorithms be made fairer by having these measurement strategies (Hellman, 2020), debiasing can be done by incorporating fairness constraints into the development phase (Zafar et al., 2017). Generally, a myriad of technical and statistical models aimed at reducing algorithmic biases have been employed (Baer & Kamalnath, 2017; Kordzadeh & Ghasemaghaei, 2021; Panch et al., 2019), with each AI model and their purpose requiring their relevant set of interventions These checks include regression (Kennedy, 1975), conducting a sensitivity analysis (Buchholz et al., 2020), hierarchical modeling (Katahira, 2016; Pollet et al., 2015), or considering alternative data points that be more accurate in predictions.

4.2.1. Regression, sensitivity analysis, and hierarchical modeling

In the context of AI and machine learning, **regression** refers to a statistical method used to model the relationship between variables, and can be used by AI models to detect and control for potential confounding variables that might introduce bias and reduce objectivity (Hort et al., 2024). For instance, Fintech models may include regressions that control for variables such as income, credit score, and employment history to ensure that financial decisions such as loans are not biased by irrelevant factors like race or gender. **Sensitivity analysis** systematically varies the AI model's input parameters and assumptions, helping to identify which features significantly impact the model's predictions (Tejani et al., 2024). This can reveal sources of bias, such as if the model's accuracy significantly drops for certain demographic groups. By doing so, it ensures that the model's conclusions remain consistent under different scenarios and assumptions, rather than being influenced by specific biases (Ennali & van Engers, 2020). Lastly, **hierarchical modeling** in AI involves structuring data into nested levels and analyzing these levels simultaneously. By doing so, it accounts for variations at different levels, reducing bias that might arise from ignoring these nested structures (Carter et al., 2024). For instance, hierarchical models can separate individual user behavior from group-level trends in social media data, allowing for more accurate predictions and insights. Such approaches lead to nuanced and robust AI systems that can reduce the risk of bias and improve the generalizability of predictions.

The aforementioned strategies work by helping developers and models better understand the impact of individual factors, before running tests to ensure that models and results are not influenced by aggregation biases. By identifying biases early in the development phase, these approaches reduce the likelihood of negative downstream consequences.

4.2.2. RLHF bias mitigation

RLHF interventions entail the matching of matching technical solutions to the models' intended purpose, as well as the diversification of feedback providers. As such, RLHF mitigation strategies often require technical expertise and in-depth industrial technology to execute (Xiao et al., 2024). Consequently, RLHF mitigation strategies can vary between different AI models (Kirk et al., 2023; Liu et al., 2024; Yu et al., 2024). However, aside from technical solutions, there should also be systemic procedures that form the basis of RLHF bias mitigation strategies, regardless of industry and intended use. Since the RLHF process relies on human feedback (Liu, 2023), arising biases can be mitigated by ensuring a diverse group of human feedback providers (Siththaranjan et al., 2023). This reduces the risk of AI models developing a skewed or biased algorithm (Kirk et al., 2023). Exposing models to a wider variety of demographics, cultures, and backgrounds enables it to better generalize across all contexts, as well as generate more diverse and inclusive responses.

4.2.3. Alignment datasets

Alternatively, developers may consider using **alignment datasets** that have been tested and proven to be source and flag out any biases within the algorithm. Alignment datasets are used to ensure that the model's output aligns with its intended use, as well as ethical guidelines (Ji et al., 2023). These datasets typically consist of labeled data where each example includes input data paired with the expected output or behavior that aligns with specific guidelines, such as reducing gender bias. For example, alignment datasets may include flagged and labeled examples of biased and unethical behavior so that the AI model can learn to avoid these patterns. By doing so, the model then adjusts its decision-making processes to produce outcomes consistent with its objectives as well as unbiased ethical standards. Researchers, for instance, tested GenderAlign, an alignment dataset aimed at mitigating gender bias in LLMs (Zhang et al., 2024). By learning from these pre-set examples, the AI model adjusts its decision-making processes, to ensure that its outputs are not only accurate but also aligned with the ethical and operational guidelines The tests showed that when ChatGPT (GPT-3.5) was aligned with GenderAlign, its responses were judged to be least biased in terms of gender 46.6 % of the time, in turn producing more objective outputs and responses. The use of alignment datasets can help ensure that AI models adhere to good ethical standards and principles.

4.3. User loop interventions

4.3.1. Randomness, FaiR, learn-to-rank

During this stage, communication with users after deployment may require developers to further optimize or even retrain the model. In cases of position bias, developers may consider building **randomness** into the model as a way to mitigate this issue (Deldjoo, 2024). The randomness ensures that results shown first, as well as top ranked results are constantly shuffled, overcoming biases generated by constant user interaction with repeated content. Another technical solution involves disentangling perceived item popularity from user interest, creating a more independent system that eliminates correlation between clicks and content accuracy (Wei et al., 2021; Zheng et al., 2021). Other models include **FAiR**: Fairness-centric model that Adaptively mitigates the popularity bias in both users and items for Recommendation (Liu et al., 2023), which is a two-way fairness discriminator aimed at targeting both users and items (Liu et al., 2023), thus reducing biases at both ends. This means that the model actively works to ensure that recommendations are fair and balanced, not only by correcting biases that might cause certain items to be over-represented due to ranking or popularity biases, but also by ensuring that users' diverse preferences are reflected in the recommendations they receive. By tackling biases on both sides of the recommendation process, FAiR aims to create a more inclusive and representative system that serves the needs of all users.

Unbiased **learn-to-rank** frameworks further support this process by optimizing ranking models through decoupling clicks and biases to create a more accurate estimate of content relevance (Ai et al., 2018). This process enables the AI model to better predict content for users by adjusting for biases inherent in user interactions. When combined with models like FAiR, which actively mitigates popularity bias on both the user and item sides, this approach leads to a recommendation algorithm that not only improves accuracy but also promotes fairness in content recommendations.

4.4. Policies

From a policy perspective, clear guidelines and training standards should be established and adhered to. These guidelines serve to set standards to ensure the ethical and fair development and deployment of AI models and should be incorporated by both private and governmental bodies. For instance, tech giants such as Microsoft's AI and Ethics in Engineering and Research (AETHER) (Microsoft, n.d.) and Meta's independent Oversight Board (Oversight Board, n.d.) are examples of internal AI review boards that oversee their new AI initiatives. By embedding these standards into their operations, organizations can help safeguard against potential abuses or biases in AI applications. AETHER outlines six guiding principles for all its AI projects: accountability, transparency, fairness, reliability and safety, privacy and security, and inclusiveness. These lay the foundation for its projects, from brainstorming to product development, ensuring that ethical considerations are not an afterthought but are integrated into the entire lifecycle of AI projects. Similarly, the Oversight Board was granted the ability to enact reform through policy recommendations and independent overruling power (Barrett, 2023). For instance, the Oversight Board exercised its power by overturning Facebook's automated decision by its algorithm to remove a post containing an image of a female nipple as part of a breast cancer awareness campaign (Oversight Board, 2020). It recognized that exceptions should be made for such cases, and subsequently recommended the notification of users whenever such automated decisions have been made. Such review mechanisms are crucial, as it provides an additional layer of accountability and ensures that AI systems are developed and implemented in a way that respects user rights and ethical standards.

There have also been attempts made to reduce algorithmic biases on a national governmental level, such as USA's AI Bill of Rights (The White House, n.d.), the European Union's Artificial Intelligence Act (AIA) (The EU Artificial Intelligence Act, n.d.), as well as ASEAN's Guide on AI Governance and Ethics (ASEAN, n.d.). The AI Bill of Rights and ASEAN Guide seeks to establish a set of guidelines and protections aimed at preventing discrimination and ensuring that AI systems are designed and implemented in ways that are fair, transparent, and accountable. This ensures the protection of consumers from the harmful impacts of AI, particularly in critical areas like healthcare, finance, and criminal justice, where biased algorithms can have serious consequences. For instance, it imposes restrictions on business practices reliant on AI (Turner Lee, 2018). Similarly, the AIA categories AI systems based on their risk levels, ranging from minimal to high, and imposes strict requirements on high-risk AI systems, including rigorous testing, documentation, and human oversight. The aim is to prevent AI systems from perpetuating biases or causing harm, especially in areas that significantly affect citizens' rights and safety.

Multinational organizations are also critical in establishing AI policies. For instance, there have been consolidated efforts by global authorities and corporations such as the Organization for Economic Co-operation and Development (OECD), United Nations Educational, Scientific and Cultural Organization (UNESCO), and United Nations (UN) OECD (n.d.); UNESCO, n.d.; UN (n.d.) to shape AI governance. In particular, the OECD has established its AI Principles, focuses on promoting the responsible use and development of trustworthy AI, emphasizing human-centered values, transparency, and accountability. These principles have been adopted by over 40 countries, providing a common framework that encourages the ethical development of AI technologies. Concurrently, UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2023), aims to guide member states in creating policies that ensure AI is used to promote human rights and sustainable development. This recommendation outlines ethical guidelines for AI applications, focusing on areas such as data privacy, algorithmic transparency, and the protection of cultural diversity. Lastly, the established High-Level Panel on Digital Cooperation by the UN (United Nations Digital-Cooperation-Panel, n.d.) underscores the importance of integrating AI ethics into broader discussions about global governance, ensuring that AI technologies contribute positively to society. These collective efforts are designed to achieve a comprehensive approach to AI regulation, where private, national, and global standards help prevent the misuse of AI and ensure that AI technologies are developed and deployed in a responsible manner.

4.5. Summary

The pitfalls associated with biases in AI models are not just technical issues but are intertwined with societal norms and prejudices that AI systems, in particular, chatbots can perpetuate. These biases can manifest in various forms, such as the reinforcement of stereotypical gender roles, the use of gendered language, and differential treatment of users based on perceived gender. Such biases often arise from the training data used to develop these models, which may reflect existing societal biases if not carefully curated. Moreover, the design and testing processes, if not inclusive of diverse perspectives, can embed these biases into the final product. Once deployed, biased chatbots can perpetuate and exacerbate these issues, influencing user interactions in ways that reinforce harmful stereotypes and contribute to the marginalization of certain groups.

Ultimately, the goal is to create AI chatbots that enhance human communication without reinforcing existing inequalities. This requires a concerted effort from researchers, developers, policymakers, and users. Interventions at every stage of the process often entail heavy technical expertise and frameworks often relevant only to specific AI models and their functions. This means that mitigation strategies are often convoluted yet extremely specific and tailored to their needs. To do so, future research should focus on refining bias detection, mitigation, and evaluation methods that go above and beyond general ethical recommendations.

As detailed above, interventions at the data level—such as suppression, massaging, and over-sampling—can act as a good foundational technique to reduce biases before they even propagate through the next phases of model development. However, it is important to note that these methods are not foolproof, and should work in tandem with good, diverse, and representative datasets (Kuhlman et al., 2020). For instance, developers should consider the improvement of dataset quality, both in terms of data sources and annotator pools (Neilsen et al., 2017; Prabhakaran et al., 2021), as important ways in ensuring that biases are minimized. This is especially since biases in AI systems are often exacerbated by skewed datasets and labelling practices. Additionally, though architectural interventions such as regression, sensitivity analysis, and RLHF bias mitigation can provide statistically rigorous means of identifying and mitigating biases, these techniques should also be further refined to ensure their applicability and generalizability to other large-scale AI systems (Kirk et al., 2023; Liu et al., 2024). For instance, RLHF may act as a cornerstone of AI model optimization, yet the diversity of feedback providers (Siththaranjan et al., 2023) and the opacity of reinforcement mechanisms warrant further research. Additionally, the use of alignment datasets, such as GenderAlign, demonstrates the potential for structured bias mitigation (Zhang et al., 2024), but therein lies the challenge of developing standardized, industry-wide datasets that maintain contextual accuracy without enforcing rigid technological constraints. During real-world deployment, biases can also emerge or even intensify from consumer use. Strategies such as randomized ranking, FaiR, and Learn-to-Rank frameworks can act as a to reduce reinforcement of popularity and ranking biases arising from the user feedback loop.

Policy and governance interventions, such as the corporate AI ethics boards and international regulatory efforts as mentioned above, are also becoming increasingly important in shaping the landscape of AI fairness. However, the efficacy and true effectiveness of these policies remain uncertain, as enforcement mechanisms, compliance, and accountability structures are still evolving (Zaidan & Ibrahim, 2024). While

international organizations such as the OECD and UNESCO have established broad ethical frameworks OECD (n.d.) UNESCO, n.d. UN (n.d.), the practical implementation of these principles varies significantly across different AI applications and geopolitical contexts. Future research should critically assess whether these regulatory efforts are meaningfully influencing AI development or whether stronger oversight mechanisms are necessary to ensure compliance.

Ultimately, bias mitigation in AI chatbots is not a one-time intervention but an ongoing process that requires constant evaluation and adaptation. Ideally, future AI and chatbot models should include checks and balances at every stage of development to ensure that they are not subject to any bias pitfalls. As AI chatbots become more integrated into everyday interactions, ensuring their fairness, inclusivity, and accountability will require sustained collaboration between researchers, developers, policymakers, and users. Future research must go beyond theoretical fairness metrics to develop pragmatic, scalable, and industry-adopted solutions that balance bias mitigation with model performance, usability, and real-world applicability.

CRediT authorship contribution statement

Jerlyn Q.H. Ho: Writing – review & editing, Writing – original draft, Project administration, Investigation, Conceptualization. **Andree Hartanto:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Andrew Koh:** Writing – review & editing, Validation. **Nadyanna M. Majeed:** Writing – review & editing, Supervision.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgements

This research was supported by grants awarded to Andree Hartanto by Singapore Management University through research grants from the Ministry of Education Academy Research Fund Tier 1 (22-SOSS-SMU-041).

References

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., & Malthouse, E. (2021). User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM conference on user modeling, adaptation and personalization* (pp. 119–129). https://doi.org/10.1145/3450613.3456821

Ai, Q., Mao, J., Liu, Y., & Croft, W. B. (2018). Unbiased learning to rank: Theory and practice. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2305–2306). https://doi.org/10.1145/3234944.3234980

Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. *Journal of Business Research, 144*, 201–216. https://doi.org/10.1016/j.jbusres.2022.01.083

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the fourth workshop on online abuse and harms* (pp. 184–190). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.21.

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law, 30*(1), 1–17. https://doi.org/10.1007/s10506-021-09286-4

Alkurd, R., Abualhaol, I. Y., & Yanikomeroglu, H. (2020). Personalized resource allocation in wireless networks: An ai-enabled and big data-driven multi-objective optimization. *IEEE Access, 8*, 144592–144609. https://doi.org/10.48550/arXiv.2307.03867

Altaf, B. (2019). A study of fairness and information heterogeneity in recommendation systems (doctoral dissertation). https://doi.org/10.25781/KAUST-1FRN7.

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*, Article 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Amazon. (2024). *Amazon's workforce data*. Retrieved August 5, 2024, from. In https://www.aboutamazon.com/news/workplace/our-workforce-data.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica, 23*, 77–91. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arcila, B. B. (2023). Is it a platform? Is it a search engine? It's ChatGPT! *The European Liability Regime for Large Language Models. J. Free Speech L., 3*, 455.

Aydin, Ö., & Karaarslan, E. (2023). Is ChatGPT leading generative AI? What is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems, 11*(3), 118–134. https://doi.org/10.21541/apjess.1293702

ASEAN. (n.d.). ASEAN Guide on AI Governance and Ethics. Retrieved August 5, 2024 from https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf.

Baer, T., & Kamalnath, V. (2017). Controlling machine-learning algorithms and their biases. *McKinsey Insights*. Retrieved August 5, 2024 from https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases.

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61. https://doi.org/10.1145/3209581

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AIDS, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the ACM on human-computer interaction, 5*(CSCW1), 1–34. https://doi.org/10.1145/3449148

Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—a user study. *Journal of the American Society for Information Science and Technology, 60*(1), 135–149. https://doi.org/10.1002/asi.20941

Barnes, J. C., Jorgensen, C., Beaver, K. M., Boutwell, B. B., & Wright, J. P. (2015). Arrest prevalence in a national sample of adults: The role of sex and race/ethnicity. *American Journal of Criminal Justice, 40*, 457–465. https://doi.org/10.1007/s12103-014-9273-3

Barry, C., & Lardner, M. (2011). A study of first click behaviour and user interaction on the Google SERP. In *Information systems development—business systems and services: Modeling and development* (pp. 89–99). https://doi.org/10.1007/978-1-4419-9790-6_7

Batra, R., Loeffler, T. D., Chan, H., Srinivasan, S., Cui, H., Korendovych, I. V., Nanda, V., Palmer, L. C., Solomon, L. A., Fry, H. C., & Sankaranarayanan, S. K. R. S. (2022). Machine learning overcomes human bias in the discovery of self-assembling peptides. *Nature Chemistry, 14*(12), 1427–1435. https://doi.org/10.1038/s41557-022-01055-3

Barrett, P. (2023). Meta's Oversight Board and the Need for a New Theory of Online Speech. Lawfare. Retrieved from August 11, 2024, https://www.lawfaremedia.org/article/meta-s-oversight-board-and-the-need-for-a-new-theory-of-online-speech.

Beck, C., Booth, H., El-Assady, M., & Butt, M. (2020). Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In S. Dipper, & A. Zeldes (Eds.), *Proceedings of the 14th linguistic annotation workshop* (pp. 60–73). Association for Computational Linguistics. https://aclanthology.org/2020.law-1.6.

Ben Letaifa, A. (2019). Chapter Four—SSIM and ML based QoE enhancement approach in SDN context. In A. R. Hurson (Ed.), *Advances in computers* (Vol. 114, pp. 151–196). Elsevier. https://doi.org/10.1016/bs.adcom.2019.02.004.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can Language Models Be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Berry, P., & Franks, T. J. (2010). Women in the world of corporate business: Looking at the glass ceiling. *Contemporary Issues In Education Research, 3*(2). https://doi.org/10.19030/cier.v3i2.171. Article 2.

Biana, H. T. (2024). Feminist Re-engineering of religion-based AI chatbots. *Philosophie, 9*(1). https://doi.org/10.3390/philosophies9010020. Article 1.

Biswas, S. (2023). Role of chat GPT in public health. *Annals of Biomedical Engineering, 51*, 3. https://doi.org/10.1007/s10439-023-03172-7

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems, 29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Borah, A., Barman, M. P., & Awekar, A. (2024). Are word embedding methods stable and should we care about it?. In *Proceedings of the 32nd ACM conference on hypertext and social media* (pp. 45–55). http://arxiv.org/abs/2104.08433.

Bosak, J., & Sczesny, S. (2011). Gender bias in leader selection? Evidence from a hiring simulation study. *Sex Roles: Journal of Research, 65*(3–4), 234–242. https://psycnet.apa.org/doi/10.1007/s11199-011-0012-7.

Bowman, N. A., Logel, C., LaCosse, J., Jarratt, L., Canning, E. A., Emerson, K. T., & Murphy, M. C. (2022). Gender representation and academic achievement among STEM-interested students in college STEM courses. *Journal of Research in Science Teaching, 59*(10), 1876–1900.

Bradley, T., & Alhajjar, E. (n.d.). AI ethics: Assessing and correcting conversational bias in machine- learning based chatbots.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information*

Processing Systems, 33, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
Buchholz, S., Gamst, M., & Pisinger, D. (2020). Sensitivity analysis of time aggregation techniques applied to capacity expansion energy system models. *Applied Energy, 269*, Article 114938. https://doi.org/10.1016/j.apenergy.2020.114938
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91). https://proceedings.mlr.press/v81/buolamwini18a.html.
Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., & Jernigan, W. (2016). GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers, 28*(6), 760–787. https://doi.org/10.1093/iwc/iwv046
Cai, Z. G., Haslett, D. A., Duan, X., Wang, S., & Pickering, M. J. (2023). Does ChatGPT resemble humans in language use? *PsyArXiv*. https://doi.org/10.31234/osf.io/s49qv [Preprint].
Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information, 13*(1), 41. https://doi.org/10.3390/info13010041
Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops* (pp. 13–18). https://doi.org/10.1109/ICDMW.2009.83
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science (New York, N.Y.), 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230
Carter, J. D., Chacón-Montalván, E., & Leeson, A. (2024). Bias correction of climate models using a bayesian hierarchical model. *EGUsphere*, 1–36. https://doi.org/10.5194/egusphere-2023-2536, 2024.
Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications, 10*(1), 1–12. https://doi.org/10.1057/s41599-023-02079-x
Chinta, S. V., Wang, Z., Zhang, X., Viet, T. D., Kashif, A., Smith, M. A., & Zhang, W. (2024). *AI-driven healthcare: A Survey on ensuring Fairness and mitigating bias* (arXiv: 2407.19655). *arXiv*. http://arxiv.org/abs/2407.19655.
Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research, 25*(1), Article e47184. https://doi.org/10.2196/47184
Chowdhary, K. R. (2020). Natural Language processing. In *Fundamentals of artificial intelligence*. New Delhi: Springer. https://doi.org/10.1007/978-81-322-3972-7_19.
Chu, C., Donato-Woodger, S., Khan, S. S., Shi, T., Leslie, K., Abbasgholizadeh-Rahimi, S., Nyrup, R., & Grenier, A. (2024). Strategies to mitigate age-related bias in machine learning: Scoping review. *JMIR aging, 7*, Article e53564. https://doi.org/10.2196/53564
Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Npj Digital Medicine, 3*(1), 1–11. https://doi.org/10.1038/s41746-020-0288-5
Collins, A., Tkaczyk, D., Aizawa, A., & Beel, J. (2018). Position bias in recommender systems for digital libraries, 335–344 https://doi.org/10.1007/978-3-319-78105-1_37.
d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data, 5*(2), 120–134. https://doi.org/10.1089/big.2016.0048
Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA dermatology, 157*(11), 1362–1369. https://doi.org/10.1001/jamadermatol.2021.3129
Dastin, J. (2018). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. In (1st ed.,, *1. Ethics of Data and Analytics* (pp. 296–299). CRC Press https://doi.org/10.1201/9781003278290-44.
De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health. *Frontiers in Public Health, 11*. https://www.frontiersin.org/articles/10.3389/fpubh.2023.1166120.
De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3287560.3287572
Deldjoo, Y. (2024). Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems*. https://doi.org/10.48550/arXiv.2401.10545
Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2020). Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 conference on empirical methods in Natural Language Processing (EMNLP)* (pp. 8173–8188). https://doi.org/10.48550/arXiv.1911.03842
Due Billing, Y., & Alvesson, M. (2000). Questioning the notion of feminine leadership: A critical perspective on the gender labelling of leadership. *Gender, Work and Organization, 7*(3), 144–157. https://doi.org/10.1111/1468-0432.00103
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
Ennali, Y., & van Engers, T. (2020). *Data-driven AI development: An integrated and iterative bias mitigation approach*.
Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender bias in chatbot design. In *In chatbot research and design* (pp. 79–93). Third International Workshop. https://doi.org/10.1007/978-3-030-39540-7_6. *CONVERSATIONS*.
Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., & Chen, H. (2023). Investigating code generation performance of ChatGPT with crowdsourcing social data. In *2023 IEEE 47th annual computers, software, and applications conference (COMPSAC)* (pp. 876–885). https://doi.org/10.1109/COMPSAC57700.2023.00117
Ganz, M., Holm, S. H., & Feragen, A. (2021). Assessing bias in medical ai. In *In Workshop on interpretable ML in Healthcare at international Connference on machine learning (ICML)*.
Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225–234). https://doi.org/10.1145/3303772.3303791
George, A. S., George, A. s, & Martin, A. (2023). A review of ChatGPT AI's impact on several business sectors. In *Partners universal international innovation journal (PUIIJ)* (Vol. 1, pp. 9–23). https://doi.org/10.5281/zenodo.7644359
Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in Natural Language understanding datasets, 1161-1166. 10.18653/v1/D19-1107 http://arxiv.org/abs/1908.07898.
Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. *Proceedings of the 2013 workshop on automated knowledge base construction*. https://doi.org/10.1145/2509558.2509563
Google Cloud. (n.d). What is Supervised Learning? | Google Clould Retrieved August 5, 2024, from https://cloud.google.com/discover/what-is-supervised-learning#how-does-supervised-learning-work.
Goryunova, E., & Madsen, S. R. (2024). Chapter 1: The current status of women leaders worldwide. https://www.elgaronline.com/edcollchap/book/9781035306893/book-part-9781035306893-10.xml.
Gutierrez, M. (2021). Algorithmic gender bias and audiovisual data: A research agenda. *International Journal of Communication, 15*, 439–461.
Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2*(4), Article 100089. https://doi.org/10.1016/j.tbench.2023.100089
Halpern, D. F., & Wai, J. (2020). Sex differences in intelligence. In R. J. Sternberg (Ed.), *The cambridge handbook of intelligence* (pp. 317–345). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108770422.015.
Hamberg, K. (2008). Gender bias in medicine. *Women's Health, 4*(3), 237–243. https://doi.org/10.2217/17455057.4.3.237
Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field. *Big Data and Cognitive Computing, 7*(2). https://doi.org/10.3390/bdcc7020062. Article 2.
Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review, 106*(4), 811–866. https://www.jstor.org/stable/27074708.
Heylighen, F. (2002). Complexity and information overload in society: Why increasing efficiency leads to decreasing control. *The Information Society, 1*(44), 11.
Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing, 1*(2), 1–52. https://doi.org/10.1145/3631326
Houser, K. A. (2019). Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stanford Technology Law Review, 22*, 290.
Hu, K. (2023). CHATGPT sets record for fastest-growing user base - analyst note *reuters*. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
Huallpa, J. J., Arocutipa, J. P. F., Panduro, W. D., Huete, L. C., Limo, F. A. F., Herrera, E. E., Callacna, R. A. A., Flores, A. A., Romero, M. Á. M., Quispe, I. M., & Hernández, F. A.(2023). Exploring the ethical considerations of using Chat GPT in university education. Periodicals of Engineering and Natural Sciences 11, 105–115. doi: 10.21533/pen.v11i4.3770.
2024 IBM. (2024). What are word embeddings?. Retrieved August 5, 2024, from https://www.ibm.com/topics/word-embeddings.
Idowu, J. A., Koshiyama, A. S., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. *Computers & Education: Artificial Intelligence, 7*, Article 100267. https://doi.org/10.1016/j.caeai.2024.100267
Jebali, M. S., Valanzano, A., Murugesan, M., Veneri, G., & De Magistris, G. (2024). Leveraging the regularizing effect of mixing industrial and open source data to prevent overfitting of LLM fine tuning. In *In International joint conference on artificial intelligence 2024 workshop on AI governance: Alignment, morality, and law*.
Jespersen, B. (2003). Why the tuple theory of structured propositions isn't a theory of structured propositions. *Philosophia, 31*, 171–183. https://doi.org/10.1007/BF02380932
Ji, K., Chen, J., Gao, A., Xie, W., Wan, X., & Wang, B. (2024). LLMs could autonomously learn without external supervision. https://doi.org/10.48550/arXiv.2406.00606.
Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2023). AI alignment: A comprehensive survey. https://doi.org/10.48550/arXiv.2310.19852.
Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: A brief primer. *Behavior Therapy, 51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Johri, S. (2023). The making of ChatGPT: From data to dialogue. *Science News*. Retrieved August 5, 2024 from https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/.

Kaiksow, F. A., & Carter, J. (2022). Barriers to representation of underrepresented and excluded populations in clinical research. In *Improving representation in clinical trials and research: Building research equity for women and underrepresented groups*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK584407/.

Kamiran, F., & Calders, T. (2012). Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems, 33*. https://doi.org/10.1007/s10115-011-0463-8

Kaplan, D. M., Palitsky, R., Arconada Alvarez, S. J., Pozzo, N. S., Greenleaf, M. N., Atkinson, C. A., & Lam, W. A. (2024). What's in a name? Experimental evidence of gender bias in recommendation letters generated by ChatGPT. *Journal of Medical Internet Research, 26*, Article e51837. https://doi.org/10.2196/51837

Karaboga, U., & Vardarlier, P. (2021). Examining the use of artificial intelligence in recruitment processes. *Bussecon Review of Social Sciences (2687-2285), 2*, 1–17. https://doi.org/10.36096/brss.v2i4.234

Karalis, V. D. (2024). The integration of artificial intelligence into clinical practice. *Applied Biosciences, 3*(1). https://doi.org/10.3390/applbiosci3010002. Article 1.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology, 73*, 37–58. https://doi.org/10.1016/j.jmp.2016.03.007

Kennedy, J. O. S. (1975). Using regression analysis to reduce aggregation bias in linear programming supply models. *Australian Journal of Agricultural Economics, 19*(1), 1–11. https://doi.org/10.1111/j.1467-8489.1975.tb00141.x

Khan, S. (2023). The ethical imperative: Addressing bias and discrimination in AI-driven education. *Social Sciences Spectrum, 2*(1), 89–96.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2023). *Understanding the Effects of RLHF on LLM Generalisation and Diversity* (arXiv:2310.06452). *arXiv*. http://arxiv.org/abs/2310.06452.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., … Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion, 99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kodiyan, A. A. (2019). An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint*, 1–19.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*, 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. https://doi.org/10.48550/arXiv.2002.11836.

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing, 14*(7), 8459–8486. https://link.springer.com/article/10.1007%2Fs12652-021-03612-z.

Landau, U. (1978). Aggregate prediction with disaggregate models: The aggregation bias behavior. *Transportation Research Record, 673*. http://onlinepubs.trb.org/Onlinepubs/trr/1978/673/673-016.pdf.

Langenkamp, M., Costa, A., & Cheung, C. (2020). Hiring fairly in the age of algorithms. https://dx.doi.org/10.2139/ssrn.3723046.

Leavy, S. (2020). Uncovering gender bias in media coverage of politicians with machine learning. https://doi.org/10.48550/arXiv.2005.07734.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

Lee, M. (2023). Christians are asking ChatGPT about god. *Is This Different From Googling? ChristianityToday.Com*. from https://www.christianitytoday.com/ct/2023/may-web-only/chatgpt-google-bible-theology-artificial-intelligence-truth.html. (Accessed 5 August 2024).

Lerman, K., & Hogg, T. (2014). Leveraging position bias to improve peer recommendation. *PLoS One, 9*(6), Article e98914. https://doi.org/10.1371/journal.pone.0098914

Li, J., Dada, A., Kleesiek, J., & Egger, J. (2023). ChatGPT in healthcare: A taxonomy and systematic review. *Healthcare Informatics*. https://doi.org/10.1101/2023.03.30.23287899 [Preprint].

Li, A. J., Krishna, S., & Lakkaraju, H. (2024). More RLHF, more trust? On the impact of human preference alignment on language model trustworthiness. https://doi.org/10.48550/arXiv.2404.18870.

Limna, P., Kraiwanit, T., Jangjarat, K., Klayklung, P., & Chocksathaporn, P. (2023). The use of ChatGPT in the digital era: Perspectives on chatbot implementation. In *Journal of applied learning and teaching* (Vol. 6, pp. 64–74). https://doi.org/10.37074/jalt.2023.6.1.32, 1.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Liu, G. K. M. (2023). *Transforming human interactions with AI via reinforcement learning with human feedback (RLHF)*. Massachusetts Institute of Technology.

Liu, Z., Fang, Y., & Wu, M. (2023). Mitigating popularity bias for users and items with fairness-centric adaptive recommendation. *ACM Transactions on Information Systems, 41*(3), 1–27. https://doi.org/10.1145/3564286

Liu, Z., Lu, M., Zhang, S., Liu, B., Guo, H., Yang, Y., Blanchet, J., & Wang, Z. (2024). Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. https://doi.org/10.48550/arXiv.2405.16436.

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. https://doi.org/10.48550/arXiv.1809.07842.

Lum, K., Zhang, Y., & Bower, A. (2022). De-biasing "bias" measurement. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3531146.3533105

Lund, B. D., & Wang, T. (2023). *Chatting about ChatGPT: how may AI and GPT impact academia and libraries?Library Hi Tech News, 40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009.

Manoharan, A. (2024). Enhancing audience engagement through ai-powered social media automation. *World Journal of Advanced Engineering Technology and Sciences, 11*(2), 150–157. https://doi.org/10.30574/wjaets.2024.11.2.0084

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 conference on fairness, accountability, and transparency*. https://doi.org/10.1145/3351095.3372867

Marinucci, L., Mazzuca, C., & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: State of the art and challenges with a focus on gender. *AI & Society, 38*(2), 747–761. https://doi.org/10.1007/s00146-022-01474-3

Martínez, N., Agudo, U., & Matute, H. (2022). Human cognitive biases present in Artificial Intelligence. *International Journal on Basque Studies, RIEV, 67*(2), 51–60.

Maserejian, N. N., Link, C. L., Lutfey, K. L., Marceau, L. D., & McKinlay, J. B. (2009). Disparities in physicians' interpretations of heart disease symptoms by patient gender: Results of a video vignette factorial experiment. *Journal of Women's Health, 18*(10), 1661–1667. https://doi.org/10.1089/jwh.2008.1007

Masiero, S., & Aaltonen, A. (2020). Gender bias in information systems research: A literature review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3751440

Mathur, V., Stavrakas, Y., & Singh, S. (2016). Intelligence analysis of tay twitter bot. In *2016 2nd international conference on contemporary computing and informatics (IC3I)* (pp. 231–236). IEEE. https://doi.org/10.1109/IC3I.2016.7917966.

Mauro, G., & Schellmann, H. (2023). 'There is no standard': Investigation finds AI algorithms objectify women's bodies | Artificial intelligence (AI. *The Guardian*. from https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies. (Accessed 5 August 2024).

Meyers-Levy, J., & Loken, B. (2015). Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology, 25*(1), 129–149. https://psycnet.apa.org/doi/10.1016/j.jcps.2014.06.003.

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4354422

Microsoft. (n.d.) Responsible AI Principles and Approach Microsoft AI. Retrieved August 5, 2024, from https://www.microsoft.com/en-us/ai/principles-and-approach.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International conference on learning representations*. https://arxiv.org/pdf/1301.3781.

Mohanani, R., Salman, I., Turhan, B., Rodríguez, P., & Ralph, P. (2018). Cognitive biases in software engineering: A systematic mapping study. *IEEE Transactions on Software Engineering, 46*(12), 1318–1339. https://doi.org/10.1109/TSE.2018.2877759

Mutasa, S., Sun, S., & Ha, R. (2020). Understanding artificial intelligence based radiology studies: What is overfitting? *Clinical Imaging, 65*, 96–99. https://doi.org/10.1016/j.clinimag.2020.04.025

Naci, H., & Soumerai, S. B. (2016). History bias, study design, and the unfulfilled promise of pay-for-performance policies in health care. *Preventing Chronic Disease, 13*, E82. https://doi.org/10.5888/pcd13.160133

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. *ACIS 2020 Proceedings*, 1–12.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, (4), 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Nasution, A. H., & Onan, A. (2024). ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. *IEEE Access, 12, 71876–71900*. https://doi.org/10.1109/ACCESS.2024.3402809

National Audit Office. (2019). *Investigation into the response to cheating in English language tests*. National Audit Office.

Navidi, N., & Landry, R., Jr. (2021). New approach in human-AI interaction by reinforcement-imitation learning. *Applied Sciences, 11*(7), 3068. https://doi.org/10.3390/app11073068

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication, 10*, 17. Retrieved from https://ijoc.org/index.php/ijoc/article/view/6277/1804.

Neville, J., Gallagher, B., Eliassi-Rad, T., & Wang, T. (2012). Correcting evaluation bias of relational classifiers with network cross validation. *Knowledge and Information Systems, 30*, 31–55. https://doi.org/10.1007/s10115-010-0373-1

Newstead, T., Eager, B., & Wilson, S. (2023). How AI can perpetuate–or help mitigate–gender bias in leadership. *Organizational Dynamics, 52*(4), Article 100998. https://doi.org/10.1016/j.orgdyn.2023.100998

Nooreyezdan, N. (2023). *India's religious AI chatbots are speaking in the voice of god—and condoning violence*. Rest of World. Retrieved August 5, 2024 https://restofworld.org/2023/chatgpt-religious-chatbots-india-gitagpt-krishna/.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns, 2*(10), Article 100347. https://doi.org/10.1016/j.patter.2021.100347

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *WIREs Data Mining and Knowledge Discovery, 10*(3), Article e1356. https://doi.org/10.1002/widm.1356

Nuseir, M. T., Al Kurdi, Alshurideh, M. T., & Alzoubi, H. M. (2021). Gender discrimination at workplace: Do artificial intelligence (AI) and machine learning (ML) have opinions about it. In *The international conference on artificial intelligence and computer vision* (pp. 301–316). https://doi.org/10.1007/978-3-030-76346-6_28.

Obermeyer, Z., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. *Proceedings of the Conference on fairness, accountability, and transparency, 89*. https://doi.org/10.1145/3287560.3287593

Obradovic, Z., & Vucetic, S. (2004). *Challenges in scientific data mining: Heterogeneous, biased, and large samples. Center for Information Science and Technology*. Temple University.

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & Society, 39*(4), 1–13. https://doi.org/10.1007/s00146-023-01675-4

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). Linguistic ambiguity analysis in ChatGPT. http://arxiv.org/abs/2302.06426.

OECD (n.d.) AI Principles Overview. Retrieved August 5, 2024, from https://oecd.ai/en/ai-principles.

Ovchinnikov, A. (2020). Ethics and ai: The 2020 international baccalaureate grading scandal (SSRN Scholarly Paper 3740517) https://papers.ssrn.com/abstract=3740517.

Oversight Board. (n.d.). Improving how Meta treats people and communities around the world. Retrieved August 5, 2024, fromn. d https://www.oversightboard.com/.

Oversight Board. (2020). Breast cancer symptoms and nudity. Retrieved August 5, 2024, from https://www.oversightboard.com/decision/IG-7THR3SI1/.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of global health, 9*(2), Article 010318. https://doi.org/10.7189/jogh.09.020318

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery, 126*(2), 619–625. https://doi.org/10.1097/PRS.0b013e3181de24bc

Perez, C.P., (2020) We Need to Close the Gender Data Gap By Including Women in Our Algorithms. Time Magazine. Retrieved August 5, 2024, from https://time.com/collection/davos-2020/5764698/gender-data-gap/.

Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *American Journal of Primatology, 77*(7), 727–740. https://doi.org/10.1002/ajp.22405

Prabhakaran, V., Davani, A. M., & Diaz, M. (2021). On releasing annotator-level labels and information in datasets. https://doi.org/10.48550/arXiv.2110.05699.

Qraitem, M., Saenko, K., & Plummer, B. A. (2023). Bias mimicking: A simple sampling approach for bias mitigation. In *In proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20311–20320).

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 118*(26). https://doi.org/10.1073/pnas.2024292118. Article e2024292118.

Restrepo Amariles, D., & Baquero, P. M. (2023). Promises and limits of law for a human-centric artificial intelligence. *Computer Law & Security Report, 48*, Article 105795. https://doi.org/10.1016/j.clsr.2023.105795

Rice, L., & Barth, J. M. (2016). Hiring decisions: The effect of evaluator gender and gender stereotype characteristics on the evaluation of job applicants. *Gender Issues, 33*, 1–21. https://doi.org/10.1007/s12147-015-9143-4

Richardson, R., Schultz, J. M., & Crawford, K. (2019). *Dirty data, bad predictions: HOW civil rights violations impact police data, predictive policing systems, and justice* (Vol. 94). NEW YORK UNIVERSITY LAW REVIEW.

Riedl, R., Hubert, M., & Kenning, P. (2011). Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers. *MIS Quarterly, 35*(1). https://doi.org/10.2307/20721434, 257–257.

Riyazahmed, K. (2023). AI in finance: Needs attention to bias. *ANNUAL RESEARCH JOURNAL OF SCMS, PUNE, 11*, 1.

Roumeliotis, K. I., & Tselikas, N. D. (2023). Chatgpt and open-ai models: A preliminary review. *Future Internet, 15*(6), 192. https://doi.org/10.3390/fi15060192

Ruiz-Gutierrez, V., Hooten, M. B., & Campbell Grant, E. H. (2016). Uncertainty in biological monitoring: A framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution, 7*(8), 900–909. https://doi.org/10.1111/2041-210X.12542

Schwab, K. (2024). *The fourth industrial revolution-what it means and how to respond*. World Economic Forum. from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/. (Accessed 5 August 2024).

Schwartz, R., Down, L., Jonas, A., & Tabassi, E. (2021). *A proposal for identifying and managing bias in artificial intelligence* (Vol. 1270). Draft NIST Special Publication. https://doi.org/10.6028/NIST.SP.1270-draft

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine, 27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2022). A survey on techniques for identifying and resolving representation bias in data. https://doi.org/10.48550/arXiv.2203.11852.

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys, 55*(13s), 1–39. https://doi.org/10.1145/3588433

Silberg, J., & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute, 1*(6), 1–31.

Singh, S., & Ramakrishnan, N. (2023). Is ChatGPT biased? A review. https://doi.org/10.31219/osf.io/9xkbu.

Siththaranjan, A., Laidlaw, C., & Hadfield-Menell, D. (2023). Understanding hidden context in preference learning: Consequences for RLHF. In *Socially responsible language modelling research*.

Stack Overflow. (2022). Stack Overflow Developer Survey, Retrieved August 5, 2024, from. https://survey.stackoverflow.co/2022/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2022.

Sun, X., Ma, R., Zhao, X., Li, Z., Lindqvist, J., Ali, A. E., & Bosch, J. A. (2024). Trusting the search: Unraveling human trust in health information from Google and ChatGPT. https://doi.org/10.48550/arXiv.2403.09987.

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics, 44*(5), Article e230067. https://doi.org/10.1148/rg.230067

Temsah, M. H., Aljamaan, F., Malki, K. H., Alhasan, K., Altamimi, I., Aljarbou, R., Bazuhair, F., Alsubaihin, A., Abdulmajeed, N., Alshahrani, F. S., Temsah, R., Alshahrani, T., Al-Eyadhy, L., Alkhateeb, S. M., Saddik, B., Halwani, R., Jamal, A., Al-Tawfiq, J. A., & Al-Eyadhy, A. (2023). ChatGPT and the future of digital health: A study on healthcare workers' perceptions and expectations. *Health Care, 11*(13), 1812. https://doi.org/10.3390/healthcare11131812

Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society, 16*(3), 252–260. https://doi.org/10.1108/JICES-06-2018-0056

The White House. (n.d.).Blueprint for an AI Bill of Rights OSTP. Retrieved August 11, 2024, from https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

The EU Artificial Intelligence Act (n.d.). The EU Artificial Intelligence Act. Retrieved August 5, 2024, from https://artificialintelligenceact.eu/.

UNESCO. (2023). Recommendation on the ethics of artificial intelligence. from https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence. (Accessed 5 August 2024).

UNESCO. (n.d.) Ethics of Artificial Intelligence, UNESCO. Retrieved August 5, 2024, from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

United Nations. (n.d.). United Nations AI Advisory Body. Retrieved August 5, 2024, from https://www.un.org/en/ai-advisory-body.

United Nations Digital-Cooperation-Panel. (n.d.). Secretary-General's High-level Panel on Digital Cooperation. Retrieved August 5, 2024, from https://www.un.org/en/sg-digital-cooperation-panel.

Van Erkel, P. F., & Thijssen, P. (2016). The first one wins: Distilling the primacy effect. *Electoral Studies, 44*, 245–254. https://doi.org/10.1016/j.electstud.2016.09.002

Varona, D., & Suárez, J. L. (2022). Discrimination, bias, fairness, and trustworthy AI. *Applied Sciences, 12*(12), 5826. https://doi.org/10.3390/app12125826

Vázquez-Ingelmo, A., García-Peñalvo, F. J., & Therón, R. (2020). Aggregation bias: A proposal to raise awareness regarding inclusion in visual analytics. In *World conference on information systems and technologies* (pp. 409–417). https://doi.org/10.1007/978-3-030-45697-9_40

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports, 13*(1), Article 15737. https://doi.org/10.1038/s41598-023-42384-8

Victor, D. (2016). Microsoft created a twitter bot to learn from users. It quickly became a racist jerk. *New York*. Retrieved August 5, 2024, from https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *Verge*. Retrieved August 5, 2024, from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Walsh, T. (2022). Everyone's having a field day with ChatGPT – but nobody knows how it actually works. *The Conversation*. Retrieved August 5, 2024, from http://theconversation.com/everyones-having-a-field-day-with-chatgpt-but-nobody-knows-how-it-actually-works-196378.

Wan, Y., Wang, W., He, P., Gu, J., Bai, H., & Lyu, M. R. (2023). Biasasker: Measuring the bias in conversational ai system. In *Proceedings of the 31st ACM joint European software engineering conference and symposium on the foundations of software engineering* (pp. 515–527). https://doi.org/10.1145/3611643.3616310

Waseem, Z., Lulz, S., Bingel, J., & Augenstein, I. (2021). Disembodied machine learning: On the illusion of objectivity in NLP. https://doi.org/10.48550/arXiv.2101.11974.

Wei, T., Feng, F., Chen, J., Wu, Z., Yi, J., & He, X. (2021). Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 1791–1800). https://doi.org/10.1145/3447548.3467289

West, M., Kraut, R., & Chew, H. E. (2019). The rise of gendered AI and its troubling repercussions. I'd blush if I could: Closing gender divides in digital skills through education. https://hdl.handle.net/20.500.12799/6598.

West, S. M., Whittaker, M., & Crawford, K. (2019). *Discriminating systems* (pp. 1–33). AI Now.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI now report 2018*. New York: AI Now Institute at New York University.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*(1), 327–350. https://doi.org/10.1146/annurev.so.18.080192.001551

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: Comments on microsoft's tay" experiment," and wider implications. *ACM SIGCAS - Computers and Society, 47*(3), 54–64. https://doi.org/10.1145/3144592.3144598

Woodall, A., Morgan, C., Sloan, C., & Howard, L. (2010). Barriers to participation in mental health research: Are there specific gender, ethnicity and age related barriers? *BMC Psychiatry, 10*, 1–10. https://doi.org/10.1186/1471-244X-10-103

World Economic Forum. (2019). *Global Gender Gap report 2020*. Retrieved 5 August 2024 from https://www.weforum.org/publications/gender-gap-2020-report-100-years-pay-equality/.

Worth, P. J. (2023). Word embeddings and semantic spaces in natural language processing. *International Journal of Intelligence Science, 13*(1), 1–21. https://doi.org/10.4236/ijis.2023.131001

Wright, W. (2024). God chatbots offer spiritual insights on demand. What could go wrong. *Scientific American*. Retrieved from https://www.scientificamerican.com/article/the-god-chatbots-changing-religious-inquiry/. (Accessed 5 August 2024).

Wu, X., Duan, R., & Ni, J. (2023). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence*. https://doi.org/10.1016/j.jiixd.2023.10.007

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems, 135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xiang, A. (2024). Mirror, mirror, on the wall, who's the fairest of them all? *Dædalus, 153*(1), 250–267. https://doi.org/10.1162/daed_a_02058

Xiao, J., Li, Z., Xie, X., Getzen, E., Fang, C., Long, Q., & Su, W. J. (2024). On the algorithmic bias of aligning Large Language Models with RLHF: Preference collapse and matching regularization. https://doi.org/10.48550/arXiv.2405.16455.

Yarger, L., Cobb Payton, F., & Neupane, B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review, 44*(2), 383–395. https://doi.org/10.1108/OIR-10-2018-0334

Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168), Article 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., & Chua, T.-S. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 13807–13816). https://doi.org/10.48550/arXiv.2312.00849

Yudkowsky, E. (2001). *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. San Francisco, USA: The Singularity Institute.

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *The Lancet Digital Health, 6*(1), e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics* (pp. 962–970). https://doi.org/10.48550/arXiv.1507.05259

Zaidan, E., & Ibrahim, I. A. (2024). AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications, 11*(1), 1–18.

Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., & Zhang, Y. (2021). Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 11–20). https://doi.org/10.1145/3404835.3462875

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: Quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM conference on health, inference, and learning*. https://doi.org/10.1145/3368555.3384448

Zhang, T., Zeng, Z., Xiao, Y., Zhuang, H., Chen, C., Foulds, J., & Pan, S. (2024). GenderAlign: An alignment dataset for mitigating gender bias in Large Language Models. https://doi.org/10.48550/arXiv.2406.13925.

Zheng, Y., Gao, C., Li, X., He, X., Li, Y., & Jin, D. (2021). Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the web conference 2021* (pp. 2980–2991). https://dl.acm.org/doi/10.1145/3442381.3449788.

Zhou, Y., Kantarcioglu, M., & Clifton, C. (2021). Improving fairness of AI systems with lossless de-biasing. https://doi.org/10.48550/arXiv.2105.04534.

Zhu, Z., He, Y., Zhao, X., & Caverlee, J. (2021). Popularity bias in dynamic recommendation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 2439–2449). https://doi.org/10.1145/3447548.3467376

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. https://doi.org/10.48550/arXiv.2301.12867.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., … Irving, G.. Fine-tuning Language Models from human preferences. https://doi.org/10.48550/arXiv.1909.08593.

Zinn, S., Landrock, U., & Gnambs, T. (2021). Web-based and mixed-mode cognitive large-scale assessments in higher education: An evaluation of selection bias, measurement bias, and prediction bias. *Behavior Research Methods, 53*(3), 1202–1217. https://psycnet.apa.org/doi/10.3758/s13428-020-01480-7.