RESEARCH & INNOVATION DIVISION

Gender Biases in Generative AI: Unveiling Prejudices and Prospects in the Age of ChatGPT

Noor ul Ain
Lahore College for Women University Lahore
***Correspondence**: noorulain1114@gmail.com

Citation | Ain. N, "Gender Biases in Generative AI: Unveiling Prejudices and Prospects in the Age of ChatGPT", MCCSS, Vol. 2, Issue. 2, pp. 85-99, June 2023
Received | May 04, 2023; **Revised |** May 24, 2023; **Accepted |** June 10, 2023; **Published |** June 15, 2023.

The advent of Advanced Natural Language Processing Models and generative AI, exemplified as ChatGPT, has exerted a significant impact on individuals' personal lives and professional endeavors since its introduction, with more expansion anticipated in the future. This study delves into the intricate landscape of gender biases entrenched within Generative Artificial Intelligence (AI), specifically focusing on the prevalence and implications within the domain of ChatGPT. Through an extensive exploration of ChatGPT's responses and interactions, this research sheds light on the nuanced manifestations of gender stereotypes, disparities, and their multifaceted impact on societal constructs. Highlighting instances of bias across various prompts, including professional scenarios, parental anecdotes, and skill prioritization in CVs, the study delineates the perpetuation of gendered notions in AI-generated content. Moreover, it underscores the perlocutionary effects of such biases, elucidating their potential to reinforce societal disparities. The research underscores the significance of ethical frameworks and regulatory measures to counteract these biases, emphasizing the pivotal role of AI in promoting transformative change and fostering gender equality. Ultimately, this inquiry advocates for an informed and proactive approach to harness the promising potential of AI while mitigating gender prejudices in the digital age. Nevertheless, this study also proposes that artificial intelligence has the potential to address prejudices and counteract gender disparities. The present discourse is around the topic of gender prejudice in the context of ChatGPT, a prominent example of big language models utilized in generative AI. The focus lies on the examination of performativity and ethical considerations associated with AI systems.

Keywords: Artificial Intelligence, Gender Disparity, Digital Age.

Introduction:

In the ever-evolving landscape of artificial intelligence, the emergence of Advanced Natural Language Processing Models like ChatGPT heralds both promise and peril. As society embraces the potential of generative AI, a critical lens reveals an underlying issue: gender biases. This exploration delves into the intricate tapestry of biases entrenched within these technological marvels, unraveling the implications, ethical quandaries, and the potential for transformative change in the realm of gender prejudice within AI systems [1]. Artificial intelligence (AI) stands as a catalyst for reshaping our professional and personal realms, with generative AI emerging as a pivotal player in this transformation. Its potential to enhance productivity and bolster global economies is widely acknowledged. Projections suggest generative AI could contribute significantly to the global economy and spur GDP and productivity growth. Forecasts even speculate that by 2030, its capacity to generate diverse outputs, spanning text, code, images, and videos, might surpass human production capabilities.

In domains such as marketing and sales, generative AI holds the promise of revolutionizing creative content generation and reshaping customer interactions. Its integration into marketing strategies for copywriting and creative endeavors is evident, as seen in collaborations between technology giant Nvidia and advertising powerhouse WPP. Given the profound influence of marketing on societal perceptions, it's evident that the impact of generative AI extends deeply into shaping our society. Furthermore, AI's impact expands beyond commercial domains, infiltrating education. Generative AI's ability to customize learning experiences at various educational levels has gained recognition. It holds the potential to enhance the learning process by offering personalized information, fostering creativity, and refining digital skills to prepare learners for the evolving demands of the workplace. Within digital publishing for education, generative AI's promise of significantly reducing content creation costs across multiple mediums like text, image, audio, and video highlights its practical benefits. However, amid the widespread advantages of generative AI in business and society, vigilance is essential regarding potential drawbacks. Concerns arise over the technology's inherent risks, including issues related to intellectual property rights, output accuracy, explain ability of results, and the amplification of biases. Notably, generative AI models, creating content based on training data patterns, pose challenges in assessing bias, given the absence of a definitive "correct" output. This necessitates evaluating a spectrum of generated content to detect potential bias patterns. Particularly concerning is the impact of generated content, especially visual material, which can directly influence perceptions, perpetuate harmful stereotypes, and distort beliefs, especially when widely distributed. For instance, as educational tools increasingly integrate generative AI, they wield substantial influence over young minds and their worldview. Tailored content reflecting potential biases may inadvertently reinforce detrimental stereotypes, profoundly shaping perceptions in ways that are difficult to rectify [2]. Generative AI models, typically trained on vast internet-sourced data, pose a significant challenge due to the lack of control over data sources.

This absence of control complicates the auditing and updating of training data to mitigate potential biases. Given the diverse perspectives, cultures, and ideologies encapsulated within this data, anticipating, let alone correcting, the multitude of inadvertent biases that seep into these models becomes increasingly arduous. Compounding this challenge is the proprietary nature of several major generative AI models, often under the ownership and maintenance of private entities, shielded from public scrutiny. This lack of transparency exacerbates the issue, restricting broader academic and societal access to assess or rectify biases within these closed systems. Consequently, despite their growing prevalence across various domains, generative AI models might unwittingly perpetuate detrimental biases and stereotypes, notably within realms like marketing and education. The absence of well-defined regulations and policies in the generative AI domain amplifies these risks, leaving room for potentially harmful applications to proliferate unchecked. According to ChatGPT 2023a, individuals might be noticed transporting either a leather briefcase or a bag. AI has gained significant attention and popularity in recent years. These tools, which are capable of generating human-like text, have been widely adopted in various domains such as natural language processing and machine translation. One prominent example is OpenAI's big language model, which has demonstrated impressive capabilities in generating coherent and contextually relevant text. The utilization of generative AI tools and technologies, including Advanced Natural Language Processing Models, has sparked considerable interest and discussion in academic and research communities [3].

The pervasive influence of AI-powered conversational agents such as ChatGPT developed by OpenAI, Bard developed by Google, and Bloom developed by Big Science, has resulted in a significant impact on both the online landscape and individuals' personal and professional spheres. AI tools are currently seen as more than mere hype in the present era [4]. ChatGPT has achieved significant scalability as an artificial intelligence platform, with a remarkable user base of 100 million active users in January 2023. The utilization of web scraping techniques, coupled with the application of neural networks for text comprehension, pattern recognition, and content generation in response to prompts, has significantly expanded the realm of what is achievable in the field of artificial intelligence. This progress has been achieved through both architectural advancements and the sheer scale of data processing, as highlighted. Generative artificial intelligence (AI) holds significant potential in various domains [5]. It offers the prospect of minimizing human error, automating repetitive tasks and processes, managing vast quantities of data, expediting decision-making processes, serving as a digital assistant to individuals, efficiently executing hazardous tasks, and enhancing workflows and processes [6]. These capabilities can be leveraged in diverse contexts, irrespective of location or time constraints. However, the process of digitalization, datafication, and the integration of artificial intelligence into all aspects of society have raised several concerns and prompted significant cautionary messages. The major ethical concerns of ChatGPT are the dangers and potential damages associated with artificial intelligence (AI). These concerns are largely attributed to a lack of openness surrounding the technology [7].

The topic of interest pertains to algorithmic data gathering, encompassing concerns related to data ownership, inquiries regarding data protection, as well as the presence of data inaccuracies and biases. This study centers on the gender biases ingrained throughout big language models and their manifestation in the generative AI's answers, which can range from overt to subtle. The initial paragraph of this study has effectively demonstrated this point: as per ChatGPT [8], an economics professor is typically depicted as a male individual. In a study conducted on ChatGPT 2023b, the question on the attire of CEOs elicited a response that mostly emphasized the prominence of business suits, with blouses, dresses, and skirts being mentioned thereafter [9][10]. It is customary for nurses to be predominantly female, as they are required to ensure that their hair is firmly tied back, away from their facial area. I'm sorry, but I need more information or text from you to provide an According to ChatGPT 2023c, individuals may be obligated to eliminate excessive jewelry from their outfit, except modest earrings for females. When prompted to provide a narrative concerning the professional choices of girls and boys, ChatGPT presents the case of Lily, an individual characterized by her inclination towards creativity and artistic pursuits [11]. The individual would dedicate a considerable amount of time within her personal space, engrossed in the process of crafting aesthetically pleasing artwork that served as a manifestation of her innermost sentiments and aspirations. The individual in question exhibited promising artistic abilities, but Ethan had consistently displayed a keen interest in the fields of science and technology [12]. When inquired about the typical personality traits associated with boys, ChatGPT advises against making generalizations.

However, it does mention several commonly observed traits in boys, including physical strength, independence, assertiveness, an inclination towards technical fields, an active and adventurous nature, as well as emotional restraint. Conversely, ChatGPT suggests that girls often exhibit traits such as empathy and nurturing, effective communication and social skills, a cooperative and inclusive attitude, emotional expression, an interest in nurturing and engaging in creative activities, as well as resilience and adaptability. Regarding the characteristic attributes commonly associated with individuals who identify as non-binary, ChatGPT presents a compilation of prevalent facets that pertain solely to their gender-related experiences. These features encompass gender identity, gender expression, self-identification, preferred pronouns, gender dysphoria, as well as engagement in advocacy and activism. Likewise, narratives featuring transgender individuals often incorporate stereotypical elements. These include the obligatory portrayal of "reaching out to the LGBTQIA+ community," the depiction of a process of "coming out to their family," who initially experience confusion but ultimately accept and support their gender identity, and the desire to assist others in navigating their paths of self-discovery. These narratives tend to focus solely on personal identity concerns and gender-related experiences, neglecting the multifaceted aspects of transgender individuals' lives [13].

Gender equality is a significant objective within the framework of the United Nations' sustainability goals. The persistence of gender prejudices and inequities is observed across all contexts, impeding societal, individual, and economic advancement for all individuals. The concept of gender is not inherently predetermined at birth, but rather constructed and acquired through social and cultural influences, as argued. The concept of gender is understood to be a social construct, wherein individuals acquire knowledge and adopt behaviors and roles associated with either femininity or masculinity. Nevertheless, it is frequently observed that girls and women have societal disadvantages in various domains such as health, income, education, protection, and overall well-being during this process. Gender inequality has negative consequences for males as well. For instance, it impacts their health, well-being, and life choices [14]. Additionally, it hinders their professional advancement and personal capabilities. Individuals who identify as gender-diverse, encompassing non-binary, genderqueer, agender, bigender, transgender, or genderfluid identities, frequently encounter exclusion from both the discourse and consideration of equality.

The act of gendering is performed in a manner that involves the translation of speech, language, and discourses into social signals, practices, and realities. In accordance with the establishment of a theoretical framework for the present study, the concept of performativity posits that language possesses significant capacities to exert influential effects. Language serves a dual purpose of describing the world and facilitating transformative change by operating as a vehicle for belief and subsequent action. Language models of significant scale extract linguistic patterns, speech patterns, and various forms of communication from online sources. These models utilize this data to produce words and formulate text-based replies [15]. Consequently, they play a crucial role in shaping and exemplifying the representation and enactment of gender within the broader societal context. The present surge of artificial intelligence (AI) is unparalleled, presenting a distinctive chance to potentially disrupt the course of gender equality and challenge traditional gender norms. In the future, generative AI has the potential to either reinforce biases and injustices or actively counteract them [16]. Generative AI models, reliant on vast internet-sourced data, grapple with a critical challenge: the lack of control over these data sources. This absence of oversight hampers efforts to audit and update training data, a crucial step in mitigating potential biases. With a mosaic of perspectives, cultures, and ideologies within this data, rectifying the multitude of inadvertent biases embedded in these models proves increasingly daunting. Adding complexity, major generative AI models often operate under proprietary ownership by private entities, shielding them from public scrutiny [17]. This lack of transparency restricts broader access for academics and society to assess or rectify biases within these closed systems. Consequently, despite their widespread use in various domains, these models risk perpetuating detrimental biases and stereotypes, particularly in fields such as marketing and education. Compounding these concerns is the absence of clear regulations and policies in the generative AI domain, leaving unchecked potential for harmful applications to proliferate.

Language Models:

Advanced Natural Language Processing Models (ANLPM) encompass computerized language models employing artificial neural networks, operating via mechanical processes. These models utilize artificial intelligence algorithms to extensively analyze online sources, gather substantial data, and employ advanced deep learning techniques to comprehend, condense, synthesize, and predict new information [18]. Falling under the category of generative Artificial Intelligence systems, ANLPM, or Language Models, are tailored to generate text-based content closely resembling human language. Specifically designed to formulate responses to user inputs, these models undergo pre-training and training using comprehensive text datasets, especially in the context of ANLPM. During pre-training, certain models, such as ProtGPT2 acquire knowledge through unsupervised or self-supervised learning. In contrast, some models, like ChatGPT 3 and 4 are fine-tuned using supervision and reinforcement learning approaches [19]. During the learning process, AI networks acquire data from various sources, making it challenging to gain insight into their opaque mechanisms. Several sources can be utilized for data collection, such as Open Web Text, books, CC-news, Common Crawl, Wikipedia, social media platforms pictures, stack overflow, Pile, GitHub, PubMed, ArXiv, and even the stock exchange.

ANLPM exhibits inherent bias due to the manner in which data pertaining to them are gathered and their sources. Bias refers to a systematic tendency to mislead, attribute errors, or distort facts in a manner that favors specific groups or views, perpetuates stereotypes, or makes inaccurate conclusions based on learned patterns. Various types of biases can be observed, including demographic biases related to factors such as gender, race, or age. Cultural biases can arise from the presence of stereotypes, while linguistic biases may be influenced by the dominant language, such as English [20]. Temporal biases can be attributed to the specific period relevant to the training data. Confirmation-based biases occur when individuals actively seek out information that aligns with their existing beliefs. Lastly, ideological and political biases can manifest as a preference for specific political perspectives or ideologies (ibid.). According to ChatGPT has been observed to have a bias towards left-leaning political perspectives. There are multiple elements that can give rise to biases in Language Models. Biases are present within the training data, as the LLM tends to encounter, incorporate, and subsequently internalize these biases during its learning process. In addition, it is important to note that the algorithms employed for data processing and learning may exhibit biases. Researcher [20] conducted a study on algorithms of oppression, revealing that search engine algorithms exhibit a tendency to favor whiteness while simultaneously engaging in discriminatory practices against individuals belonging to racial minority groups, including women of color. Biases may be included in ANLPM through (semi)supervised learning, wherein human annotators offer labels and annotations to the training data based on their personal opinions, experiences, and perceptions of the world. It can be argued that the vast amount of data and information available on the internet necessitates the implementation of labels, filters, and feedback mechanisms in order to establish a coherent structure, prioritize content, and mitigate information overload. Nevertheless, it is not uncommon for technology to assume the role of "emergent gatekeepers" after biases are added through algorithmic design or human influence (ibid.). The design choices made in product development, including the design of user interfaces, have the potential to induce biases [21]. Policy decisions have a significant impact on biases, as they can shape the behaviors that tech developers enable or disable.

The prevalence of biases within AI technologies has been a long-standing issue in the market. A study conducted by the Berkeley Haas Center for Equity, Gender, and Leadership spanning 1988 to 2021 examined 133 AI systems, revealing that 44 percent exhibited gender biases and over 25 percent displayed biases related to both gender and race. Notably, this research predates the emergence of ChatGPT and Google Bard, indicating a persistent concern regarding gender biases in AI that hasn't shown improvement. One critical factor contributing to these biases is the massive volume of data integrated into Language Models (ANLPM), inevitably containing stereotypical associations and disparaging depictions of gender. The sheer volume of data within ANLPM doesn't inherently implies diversity. Instead, utilizing extensive, unfiltered training datasets often leads to the propagation of prevailing or dominant perspectives. Consequently, the training of Language Models (ANLPM) with such datasets has entrenched gender biases within these artificial intelligence systems [22]. For instance, biases can infiltrate Language Models (ANLPM) through gendered expressions, like referencing 'women doctors.' They may also manifest in negative, micro aggressive, or abusive perspectives towards genders, trivializing experiences of sexism or excluding non-binary identities (ibid.). Empirical evidence from a study point to ChatGPT reinforcing gender stereotypes. This is evident in the model assigning genders to specific professions, depicting doctors as male and nurses as female. Moreover, ChatGPT associates certain activities with genders, portraying women as primarily responsible for household chores like cooking and cleaning. Another study revealed a prevalent perception: a competent scientist is typically envisioned as a Caucasian male, while females are viewed as less capable of managing complexities in an engineering curriculum. It's been previously noted that artificial intelligence (AI) not only embodies white and male characteristics in its programming but is also visually represented in that manner [23]. The prevalence of gender biases within ChatGPT extends across various domains. The algorithm displays clear gender-based discrimination in its assessment of intelligence, as documented. Moreover, it exhibits a tendency to modify non-gendered pronouns, as highlighted in the findings. This results in the system generating gender-related disadvantages in areas like hiring, lending, and education, as observed in research. Notably, content moderation also falls prey to gender biases. Comments related to a specific gender, like women, often trigger content policy violations and feedback mechanisms, while similar content associated with the opposite gender, such as men, doesn't prompt concerns or alerts. Statista's 2023 data on ChatGPT's user base further illuminates this issue, with 65.7 percent male users and 34.3 percent female users. Consequently, there's a higher likelihood of feedback submissions from men compared to women, introducing an added layer of gender-based power dynamics and predispositions. Having explored biases in generative AI, specifically those linked to gender, the subsequent section will delve into the problematic nature of gendering and underscore the critical importance of addressing these biases [24].

Academic discourse on gender extensively explores gendering, gender bias, and the performative nature of gender. These concepts unravel the societal construction and cultural norms shaping gender roles and stereotypes, revealing how individuals are assigned and impacted by them. Gendering processes lead to biases, aligning with the notion that gender reality relies on its enactment. Early exposure to ideological influences and power dynamics, evident from birth with statements like "It's a girl," significantly shapes future experiences, often unnoticed. Gender socialization kicks off during early development, emphasizing attributes like appearance and emotional expression while sidelining traits like rationality and leadership. These gender roles persist through parenting practices, education, and into adulthood, shaping gender across politics, culture, structures, hierarchies, and behaviors, often detrimentally impacting society. Achieving gender equality, acknowledging female struggles, and addressing power imbalances present significant challenges. Gender domination, rooted in power dynamics, reinforces established norms and hierarchies, hindering progress toward equality, as highlighted by reports from the OECD and the UN. Individuals often draw conclusions based on societal gender norms, impacting perceptions across diverse contexts. In her notable work "The Psychic Life of Power," the author delves into "rapture and subjection," describing the emotional bonds women navigate while nurturing familial ties, perpetuating domestic roles and unpaid caregiving. This gender divide persists even in paid labor [25]. In "Gender Trouble," the author examines societal expectations' normalization and naturalization in practical scenarios.

Gender operates as a continuous process involving self-presentation through bodily expression. It entails repetitive actions conforming to established regulations, gradually solidifying to create an illusion of inherent naturalness. A political study of gender aims to deconstruct its tangible manifestations by analyzing and understanding the fundamental actions constituting it. It delves into the regulatory frameworks established by societal entities shaping gender norms. This construction of gender permeates everyday life including norms, cultural scripts, and social performances, resulting in prevalent stereotypes. These stereotypes contribute to specific behavioral patterns, prejudices, and biases in contemporary society. Gender bias extends beyond preferential treatment for males over females or binary identities over diverse identities. It encompasses deeply rooted concepts and biases within societal beliefs, impacting individuals across gender identities, including women, men, and gender-diverse individuals [26]. These biases can lead to an unfair distribution of resources and reinforce harmful stereotypes, perpetuating injustices that marginalize already vulnerable gender identities. Overall, gender disparities deeply influence individuals and societal spheres, embedding biases and prejudices in their personal and professional lives, and affecting various facets of societal, political, and economic realms.

Gender Performativity:

As mentioned earlier, the establishment of gender hinges on repeated societal behaviors. Gender roles and norms derive from frequent references to established societal expectations, particularly in modern times, disseminated through generative AI systems. Drawing on scholarly works, it is argued that gender norms are not simply a result of occasional verbal expressions but are shaped by ongoing discourse. The concept of performativity emphasizes how discourse iteratively shapes and confines phenomena. Additionally, it challenges the idea that outward manifestations solely represent gender's intrinsic nature. Social structuration plays a pivotal role in shaping gender within cultural contexts, influenced by societal norms and expectations. However, Butler suggests that these norms can be disrupted. Undoing gender happens when social interactions are less influenced by gender, reducing its relevance and disconnecting gendered interactions from perpetuating inequality. This transformation relies on collaborative efforts between institutions and interactions to bring about meaningful changes. AI technologies are rapidly advancing and expanding their influence across various sectors. This study asserts that generative AI, due to its broad impact, can act as a catalyst for transformative change in 'undoing gender'. This involves minimizing gender biases and promoting equality, utilizing the expansive scope and potential of generative AI.

The inquiry focuses on the intersection of generative AI and gender performativity. The cited work argues that voice and communication possess considerable effectiveness in initiating actions. Utterances, referred to as 'illocutionary acts,' can be viewed as fundamental sentences. However, language's performative nature goes beyond describing reality; it actively transforms the world it depicts, known as "illocutionary acts" when realized. Moreover, certain speech actions, termed 'perlocutionary acts,' influence recipients, yielding outcomes like persuasion, intimidation, or inspiration [20]. Unlike illocutionary acts focused on intended action, perlocutionary acts emphasize the situational impact on recipients. In this discourse, we engage in a locutionary act, expressing specific sentences with distinct meanings, akin to the classical notion of "meaning." Humans engage in illocutionary acts, like providing information or issuing orders, exerting a certain force in their speech. The study delves into illocutionary, locutionary, and perlocutionary aspects of speech acts. Locutionary elements relate to ChatGPT's responses, evidencing gendered perspectives. For instance, the AI portrayed an economics professor as male, assigning specific gender-based traits without including gender-diverse personalities. It also labeled a girl as an 'emotional artist' and a boy as a 'disruptive engineer'. ChatGPT also tends to overlook non-gendered pronouns. In a narrative prompt, 'they' and 'their' were revised to 'she' and 'her' to match the generated story's gendered identity, depicting a female protagonist named Sarah, aspiring to be a writer, not an engineer or economics professor.

Gender Equality and Societal Constructs:

The biased responses exhibited by ChatGPT and similar language models highlight their impact beyond text generation. Generative AI's reactions wield substantial influence on gender equality, potentially reinforcing biases ingrained in various text-based outputs like research articles, resumes, cover letters, essays, music, stories, conversations, recommendations, and humor. For instance, in response to a request for a narrative on a significant professional failure involving a male and female individual, ChatGPT (2023j) depicted an incident during an office dance competition. The story detailed a mishap wherein Steve's attempt to assist Lisa led to her embarrassment as her clothing tore, exposing her undergarments, emphasizing humiliation and objectification rather than fostering gender equality. Another prompt about parental abilities resulted in a biased narrative by ChatGPT (2023k). It portrayed Sarah as the nurturing mother, emphasizing her caretaking roles and creating a stereotype. In contrast, Michael was depicted as an adventurous father, engaging in interactive activities with Emily. This portrayal reinforces stereotypical gender roles, promoting emotional responsibilities for women within societal constructs, while presenting men as adventurers and educators. In analyzing the talents recommended for a 40-year-old woman and man's CVs by ChatGPT (2023l, 2023m), similarities were observed, yet subtle gender-based disparities existed. ChatGPT exhibited noticeable differences in skill prioritization between genders: technical skills ranked third for men but ninth for women. Similarly, communication and interpersonal skills ranked third for women but fifth for men.

Notably, organizational and time management skills were solely highlighted for women, whereas project management skills were exclusively mentioned for men. Following the suggested emphasis on "soft" abilities for women and "hard" skills for men in CV, revisions could inadvertently reinforce gender biases. Moreover, increased user engagement with Language Generation Models (ANLPM) may disseminate biased texts further, potentially amplified by future autonomous AI dissemination. This escalation of gender biases might result in widespread harm. Recognizing the perlocutionary impact of AI discourse on gender is crucial. User interactions with ChatGPT involve seeking information, posing inquiries, or generating content, highlighting AI's potential to sway, persuade, or influence individuals. The influence of textual content extends to AI-generated discussions, impacting readers' perceptions and responses. A collection of illustrative ChatGPT responses showcases these influences on readers. The observation posits that Language Models with Large Memory (ANLPM) can proficiently generate fluent and coherent responses, imparting an air of informativeness, persuasiveness, and authority. ChatGPT, for instance, crafts seemingly coherent and knowledgeable content, projecting an illusion of reliability. Yet, users often encounter challenges in fact-checking or verifying the accuracy of this information, consuming seemingly authoritative replies without means to confirm their accuracy.

Research [23] suggests that while ANLPM can string words together coherently, they lack consistent precision, reliability, or substantial value in the information provided. Moreover, these models fall short of meaningful communication or acknowledging human beliefs, leading to a fundamental absence of shared ideas or mutual understanding. Despite users perceiving a dialogue with the LLM, it doesn't recognize challenges to its authorship or respond to pushback, limiting user interactions to providing positive or negative feedback. Ienca raises concerns about AI's potential for "digital manipulation," surpassing previous methods in subtlety, automation, and reach. ANLPM wields a persuasive authority that significantly influences users' professional and personal lives, potentially reinforcing gendered perspectives and undermining individual or group agency, even when bias is acknowledged. They lack reliability in providing accurate information and can engage in manipulative behaviors. The discourse on "Undoing Gender: Social Change, Ethics, and Practice" revolves around challenging and reshaping traditional gender norms and roles within society. It delves into the ethical implications and practical strategies involved in dismantling and redefining gender constructs. This study aimed to investigate the depth of gender bias within ChatGPT, exploring how gender is enacted and represented within generative AI systems. It uncovered evident gender bias in ChatGPT, supported by specific instances extracted from the platform, corroborating findings from other researchers. However, further empirical studies are essential to comprehend the scale and impact of gender biases ingrained within Language and Machine Learning Models (ANLPM) and Artificial Intelligence (AI) technologies.

This exploration should delve into how these biases influence beliefs and actions. A significant theoretical contribution of this study involves expanding on Butler's foundational work, offering a fresh perspective on gender performativity within generative AI. This examination focused on the verbal expressions, actions, and consequences of gender performativity, revealing that conversational AI while appearing authoritative, relies on flawed information and biased datasets. These AI dialogues exhibit performative traits that could perpetuate gender biases and reinforce established perspectives. There's an opportunity highlighted by research to rectify gender biases using AI as a catalyst for change. The subsequent sections offer an overview of ongoing discussions aiming to mitigate gendered perspectives and prejudices. Social change encompasses the transformation of societal structures, norms, values, and behaviors over time, where discourses such as scripts, documents, and AI outputs—both reflect and shape societal dynamics. ANLPM act as mirrors reflecting existing gender biases, inequities, and injustices present in the external environment. The intricate relationship between AI and gender issues stems from AI being a human creation, inheriting human fallibility. Comprehensive analyses underscore the complexities in addressing biases in artificial intelligence, acknowledging the absorption of biases, preconceptions, and assumptions inherent in human language, especially during unsupervised AI learning.

These models merely mirror prevailing social values, norms, and cultural behaviors, including power dynamics and political structures. Moreover, cultural norms and values can vary significantly across countries and communities, impacting perceptions of bias and fairness. Fairness, particularly in contexts involving diverse stakeholders with varied perspectives, remains subjective. AI acts as a platform converging different stakeholders, such as tech companies, investors, researchers, regulators, policymakers, and communities, presenting an array of perspectives and challenges in navigating biases and fairness. Certainly, the perspectives and goals of these stakeholders often clash rather than align. Additionally, language remains dynamic, continuously evolving and transforming. As a result, certain biases ingrained within AI systems may persist despite efforts to mitigate them. Further scholarly inquiry is crucial for a more comprehensive understanding of this complex matter. It's essential to recognize that social transformation is achievable, including the deconstruction of the concept of gender. This can be realized through various approaches, such as acknowledging the diverse challenges faced by different genders, reassessing political systems, and addressing prevailing power imbalances. Artificial intelligence (AI) serves as a valuable tool for comprehending gender biases, inequalities, and injustices in society. By studying AI's reflection of the external environment, researchers can probe the individuals or factors contributing to these gender-related issues. Moreover, AI enables an examination of the societal implications of these gender concerns and facilitates an understanding of how gendered perspectives might evolve over time.

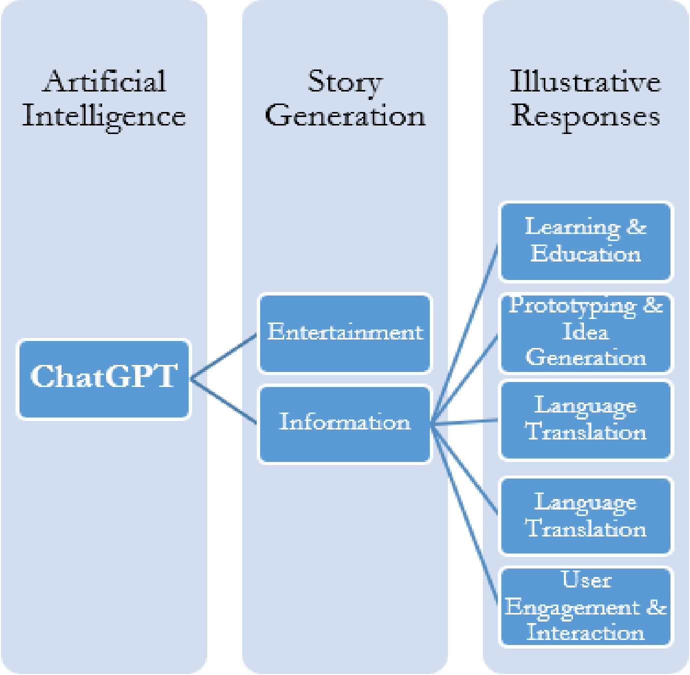

Figure 1: Kind of Responses that can be generated by ChatGPT.

Gender-Neutral Artificial Intelligence: Challenges and Principles:

AI stands as a significant catalyst for promoting gender equality across society, influencing both political discourse and practical applications. Challenging and transforming gender involves actively opposing gendered characteristics, prejudices, and stereotypes, resisting established gender recognition frameworks, and ensuring that AI's outcomes remain uninfluenced by gender biases. To enable transformative change and dismantle gender norms, it's crucial to exercise control, rectify flaws, and implement regulatory measures governing AI. Studies, including this one, highlight the potential amplification of gendered perspectives and prejudices with the widespread adoption of generative AI. Acknowledging AI's substantial future implications requires ethical considerations, regulatory frameworks, and corrective measures. Without these, AI may inadvertently inflict more harm than benefit on society. Given that gender is influenced and enacted by scripts and norms within AI systems, ethical considerations, and practical actions are essential. The focus shifts from whether AI will impact society to understanding the agents, methods, locations, and timing of its positive or negative consequences. The performative effects of AI underscore the necessity for ethical principles governing its development and expansion, aiming to offer beneficial opportunities for all. Ethical AI requires moving away from Silicon Valley's ethos of rapid progress and disruptive norms, emphasizing cautious progress and avoiding unintended repercussions. Prioritizing principles like beneficence, upholding human dignity, and enhancing well-being, alongside non-maleficence, safeguarding privacy, and ensuring security, is crucial. Autonomy in decision-making and justice, promoting societal prosperity and solidarity, further underpin ethical AI considerations.

The framework proposed offers four key principles to address biases, particularly gender-related ones, in responsible artificial intelligence (AI) research. One vital element involves representation, requiring training data to reflect a diverse range of viewpoints, experiences, and backgrounds mirroring society's heterogeneity. Emphasizing representation in AI systems is pivotal for ensuring fairness, equality, and democracy. A study by the European Institute for Gender Equality revealed that only 12% of individuals with over a decade of AI expertise are women, while women's representation in any AI role stands at 20%. Additionally, 2023 Statista data indicates that roughly two-thirds of ChatGPT users are male, suggesting the platform's primary focus on a male audience. Transparency, as proposed, involves full disclosure of methodology, data sources, and inherent constraints in AI models. Presently, Language and Machine Learning Models (ANLPM) are often referred to as "black boxes" in academic literature due to their lack of transparency. Accountability encompasses monitoring AI models, employing bias-mitigating techniques, and responsiveness to user and community concerns. Inclusivity ensures AI systems are accessible to all users, considering gender, language, culture, and individual needs. Further exploration may uncover additional foundational aspects vital for advancing ethical AI.

Ethical considerations imply that AI can rectify past harm, promote gender equality, and facilitate human self-actualization without compromising capabilities. AI empowers individuals while preserving their responsibilities, enhances social capabilities while retaining human control, and fosters societal unity without infringing on human autonomy. Subsequent phases involve strategic formulation, operational execution, responsibility, and supervision. Establishing ethical artificial intelligence (AI) stands as both a goal and a necessity. Yet, the practical steps to address biases and rectify gender disparities within AI systems demand attention. Representation, inclusivity, and transparency necessitate gender-aware methodologies to tackle biases. Four strategies can drive this endeavor. Firstly, technology firms must fortify gender diversity, equity, and inclusion within their development teams, showcasing a tangible commitment to these values. More diverse human agents, representing varied perspectives, are crucial in annotating data for AI training, clarifying AI confusion, and engaging users effectively. Gender equality in leadership is equally vital. Secondly, acknowledging inherent biases in data and algorithms is pivotal. A suggested pause in AI's current state advocates against an indiscriminate gathering of vast datasets, emphasizing meticulous planning to mitigate real-world gender-related risks. Ethical frameworks should guide data selection, as emphasized by experts.

However, acknowledging AI's irreversible advancement, continual monitoring of practices and full disclosure of AI models' methodology, data sources, and potential biases become paramount. Achieving this requires reliable self-evaluation or regulatory measures. The clear articulation of technology providers' intentions, values, and goals in developing artificial intelligence (AI)—including algorithm design, training, data assembly, and moderation—is vital. Users must receive education and awareness about biases, and tech businesses should take responsibility for equipping them with resources and support to address these issues. Public awareness campaigns regarding the sociological, legal, and ethical implications of AI are crucial. Collaborating with experts in feminist data practices and digital democracy can aid in comprehensive audits, understanding gender implications in AI, and integrating gender-sensitive concepts. Prioritizing the inclusion of marginalized groups like women and non-binary individuals in AI development is essential. Participative approaches and insights from other sectors can offer strategies to address gender disparities effectively. Designating a responsible individual within tech businesses for AI ethics is recommended. However, the dissolution of Google's entire 'Ethical AI' team in 2021 underscores the substantial progress that tech corporations still need to make in embedding and sustaining ethical frameworks. Subverting traditional gender norms using generative artificial intelligence requires meticulous coordination and effective oversight. Accountability and governance demand substantial time and resources for engaging key stakeholders and establishing technology ecosystems that ensure fair benefits for all involved parties.

However, given that AI technologies are already in the market, their impact on gender justice and inequality warrants thorough evaluation. Remedying errors and resultant damages may necessitate reevaluating or establishing regulations to adapt ethical considerations to technology's swift advancements. Developing audit processes to enforce compliance and criteria for assessing AI system trustworthiness is crucial. Context-specific restrictions on AI model applications might be essential, especially for jobs or decision-making where societal harm is plausible. Entities like the European Union and the World Economic Forum's AI Governance Alliance consider these factors. Supervision for public welfare should rest with oversight authorities at various levels, ensuring monitoring systems and effective user reporting. Independent regulatory bodies dedicated to overseeing AI can foster stakeholder collaboration and offer vital recommendations. Future research should monitor the practical use of ethical AI and its potential to mitigate gender biases and inequalities in the real world.

Conclusion:

In summary, this discussion provides a thorough exploration of the subject matter. When discussing the significance of gender equality, ChatGPT (2023n) highlights its moral imperative and strategic value. Gender equality holds importance across various domains, encompassing human rights, social equity, economic benefits, health, democratic governance, societal progress, and innovation. ChatGPT emphasizes the need for collaborative efforts among individuals, communities, governments, and organizations to challenge gender stereotypes and eliminate discriminatory practices for an inclusive society (ChatGPT, n.d.). However, as evidenced in this study, ChatGPT exhibits noticeable gender bias, perpetuating existing disparities and further disadvantaging individuals across gender identities. Generative AI holds significant potential to enhance societal well-being, but this potential relies on implementing ethical and regulatory frameworks. Without such measures, these technologies won't effectively drive transformative change or contribute to achieving gender equality, despite their inherent capabilities. Yet, the conclusion remains optimistic. The development, implementation, and scalability of generative AI are open to modifications, and ongoing efforts aimed at understanding AI's societal impacts continue. While complex, it's feasible to foster positive transformation in the 21st century and shape a future aligned with gender parity goals.

References:

[1] N. Gross, Gross, and Nicole, “What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI,” Soc. Sci., vol. 12, no. 8, pp. 1–15, 2023, Accessed: Sep. 20, 2023. [Online]. Available: https://econpapers.repec.org/RePEc:gam:jscscx:v:12:y:2023:i:8:p:435-:d:1208555

[2] D. V. P. S. , “How can we manage biases in artificial intelligence systems – A systematic literature review,” Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, Apr. 2023, doi: 10.1016/j.jjimei.2023.100165.

[3] N. Gross and S. Geiger, “A Multimethod Qualitative Approach to Exploring Multisided Platform Business Models in Health Care,” A Multimethod Qual. Approach to Explor. Multisided Platf. Bus. Model. Heal. Care, Mar. 2023, doi: 10.4135/9781529670547.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” FAccT 2021 - Proc. 2021 ACM

Conf. Fairness, Accountability, Transpar., pp. 610–623, Mar. 2021, doi: 10.1145/3442188.3445922.

[5] S. L. Beilock, E. A. Gunderson, G. Ramirez, and S. C. Levine, "Female teachers' math anxiety affects girls' math achievement," Proc. Natl. Acad. Sci., vol. 107, no. 5, pp. 1860–1863, Feb. 2010, doi: 10.1073/PNAS.0910967107.

[6] M. L. Phan and D. S. Vicario, "Hemispheric differences in processing of vocalizations depend on early experience," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 5, pp. 2301–2306, Feb. 2010, doi: 10.1073/PNAS.0900091107.

[7] H. K. Kwon et al., "Generation of regulatory dendritic cells and CD4+Foxp3 + T cells by probiotics administration suppresses immune disorders," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 5, pp. 2159–2164, Feb. 2010, doi: 10.1073/PNAS.0904055107.

[8] J. R. Poganik and V. N. Gladyshev, "We need to shift the focus of aging research to aging itself," Proc. Natl. Acad. Sci. U. S. A., vol. 120, no. 37, Sep. 2023, doi: 10.1073/PNAS.2307449120.

[9] M. S. Shukla et al., "Remosomes: RSC generated non-mobilized particles with approximately 180 bp DNA loosely associated with the histone octamer," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 5, pp. 1936–1941, Feb. 2010, doi: 10.1073/PNAS.0904497107.

[10] A. Ali, "Gendered Perspectives in Workplace Dynamics: Supervisory Satisfaction and Work-Life Balance," Magna Cart. Contemp. Soc. Sci., vol. 2, no. 2, pp. 55–64, May 2023, Accessed: Jan. 06, 2024. [Online]. Available: https://journal.50sea.com/index.php/MC/article/view/643

[11] R. Dwyer, A. Palepu, C. Williams, D. Daly-Grafstein, and J. Zhao, "Unconditional cash transfers reduce homelessness," Proc. Natl. Acad. Sci., vol. 120, no. 36, Sep. 2023, doi: 10.1073/PNAS.2222103120.

[12] S. J. Lymbery, B. L. Webber, and R. K. Didham, "Complex battlefields favor strong soldiers over large armies in social animal warfare," Proc. Natl. Acad. Sci. U. S. A., vol. 120, no. 37, Sep. 2023, doi: 10.1073/PNAS.2217973120.

[13] S. L. Beilock, E. A. Gunderson, G. Ramirez, and S. C. Levine, "Reply to Plante et al.: Girls' math achievement is related to their female teachers' math anxiety," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 20, May 2010, doi: 10.1073/PNAS.1003899107.

[14] T. Breda and C. Napp, "Girls' comparative advantage in reading can largely explain the gender gap in math-related fields," Proc. Natl. Acad. Sci. U. S. A., vol. 116, no. 31, pp. 15435–15440, Jul. 2019, doi: 10.1073/PNAS.1905779116.

[15] N. T. T. Lau, Z. Hawes, P. Tremblay, and D. Ansari, "Disentangling the individual and contextual effects of math anxiety: A global perspective," Proc. Natl. Acad. Sci. U. S. A., vol. 119, no. 7, Feb. 2022, doi: 10.1073/PNAS.2115855119.

[16] E. Bozdag, "Bias in algorithmic filtering and personalization," Ethics Inf. Technol., vol. 15, no. 3, pp. 209–227, Sep. 2013, doi: 10.1007/S10676-013-9321-6/METRICS.

[17] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating contextual information in recommender systems using a multidimensional approach," ACM Trans. Inf. Syst., vol. 23, no. 1, pp. 103–145, Jan. 2005, doi: 10.1145/1055709.1055714.

[18] S. L. Althaus and D. Tewksbury, "Agenda setting and the 'new' news," Communic. Res., vol. 29, no. 2, p. 180, Apr. 2002, doi: 10.1177/0093650202029002004.

[19] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," WWW'12 - Proc. 21st Annu. Conf. World Wide Web, pp. 519–528, 2012, doi: 10.1145/2187836.2187907.

[20] J. Bar-llan, K. Keenoy, M. Levene, and E. Yaari, "Presentation bias is significant in determining user preference for search results—A user study," J. Am. Soc. Inf. Sci.

Technol., vol. 60, no. 1, pp. 135–149, Jan. 2009, doi: 10.1002/asi.20941.

[21] K. Barzilai-Nahon, "Toward a theory of network gatekeeping: A framework for exploring information control," J. Am. Soc. Inf. Sci. Technol., vol. 59, no. 9, pp. 1493–1512, Jul. 2008, doi: 10.1002/asi.20857.

[22] K. Barzilai-Nahon, "Gatekeeping: A critical review," Annu. Rev. Inf. Sci. Technol., vol. 43, no. 1, pp. 1–79, 2009, doi: 10.1002/aris.2009.1440430117.

[23] J. Van Cuilenburg, "On competition, access and diversity in media, old and new some remarks for communications policy in the information age," New Media Soc., vol. 1, no. 2, pp. 183–207, 1999, doi: 10.1177/14614449922225555.

[24] M. J. Eppler and J. Mengis, "The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines," Inf. Soc., vol. 20, no. 5, pp. 325–344, Nov. 2004, doi: 10.1080/01972240490507974.

[25] H. Garcia-Molina, G. Koutrika, and A. Parameswaran, "Information seeking," Commun. ACM, vol. 54, no. 11, p. 121, Nov. 2011, doi: 10.1145/2018396.2018423.

[26] L. A. Granka, "The politics of search: A decade retrospective," Inf. Soc., vol. 26, no. 5, pp. 364–374, 2010, doi: 10.1080/01972243.2010.511560.