APSIPA Transactions on Signal and Information Processing, 2024, 13, e102

Overview Paper

Bias and Fairness in Chatbots: An Overview

Jintang Xue[1*], Yun-Cheng Wang[1], Chengwei Wei[1], Xiaofeng Liu[2], Jonghye Woo[2] and C.-C. Jay Kuo[1]

[1] *University of Southern California, Los Angeles, California, USA*
[2] *Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114, USA*

ABSTRACT

Chatbots have been studied for more than half a century. With the rapid development of natural language processing (NLP) technologies in recent years, chatbots using large language models (LLMs) have received much attention nowadays. Compared with traditional ones, modern chatbots are more powerful and have been used in real-world applications. There are, however, bias and fairness concerns in modern chatbot design. Due to the huge amounts of training data, extremely large model sizes, and lack of interpretability, bias mitigation and fairness preservation of modern chatbots are challenging. Thus, a comprehensive overview on bias and fairness in chatbot systems is given in this paper. The history of chatbots and their categories are first reviewed. Then, bias sources and potential harms in applications are analyzed. Considerations in designing fair and unbiased chatbot systems are examined. Finally, future research directions are discussed.

Keywords: Chatbots, ChatGPT, Bias, Fairness, Natural Language Processing.

*Corresponding author: Jintang Xue, jintangx@usc.edu.

Received 15 September 2023; Revised 06 November 2023
ISSN 2048-7703; DOI 10.1561/116.00000064

1 Introduction

A chatbot is an intelligent software system designed to simulate natural human language conversations between humans and machines [35]. As a human-computer interaction (HCI) system [41], it takes human voice or text as input and uses the natural language processing (NLP) technology to understand and respond accordingly [5]. With the rapid development of the Internet and artificial intelligence (AI), chatbots have become a hot research topic and a real-world application system that attracts much attention [136].

One of the most common occasions is to use chatbots as a dialogue agent in the service industry [4, 110, 175]. Chatbots have changed the way customers and companies interact. While chatbots may not be as good as human services in answering complex questions, they are accessible, responsive, and always available. They can answer most simple questions, which proves to be valuable in applications like product ordering and travel booking [92, 117, 186]. For companies, chatbots can respond to customer requests at any time, improve user experience, and contribute to saving in the service cost [220]. As to users, a study [30] showed that people would be interested in chatbot services for effective and efficient information access. Other motivations include entertainment, socializing, and curiosity about new things. To realize these benefits, chatbots need to understand user input and analyze users' sentiments and intentions accurately, find appropriate answers, and generate fast and fluent responses. Sometimes, it may need to take the user identity (or attributes) into account in providing a proper answer.

Recent advances and breakthroughs in NLP and machine learning (ML) have changed the landscape of language understanding and processing [97, 139, 198]. These developments are driven by the availability of increased computing power, massive amounts of training data, and the advent of sophisticated ML algorithms. The introduction of transformer networks [190] leads to large pre-trained models, such as GPT-3 [31], BERT [56], PaLM [44], etc. They have become popular [93, 146, 196, 214] in the past decade. Based on these developments, ChatGPT, a chatbot from OpenAI, has taken the world by storm by providing real-time, plausible-looking responses to input questions. ChatGPT has a good performance in text generation, language understanding, and translation. As a chatbot, it can be applied in various fields [71], such as education [14, 67], healthcare [23, 158], marketing [89, 151], environmental research [22, 218], etc. The prevalence of ChatGPT has made chatbots a focus of attention. Leading technology companies have also released their own chatbots, such as Google's Bard and Meta's BlenderBot 3.

With the help of AI, chatbots have become more intelligent and can answer people's questions smoothly. On the other hand, chatbots are not as neutral as expected, raising ethical concerns among the general public [144]. Figure 1 shows the number of papers on chatbot since 2014. The number of papers

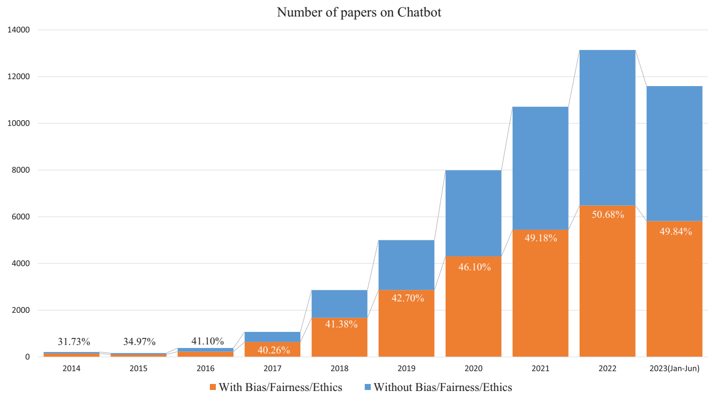

Figure 1: Search results by the year for "chatbot" or "ChatGPT" as keywords in the website of "Dimensions" since 2014.

on chatbots has risen sharply since 2015. Noticeably, the number of papers in the first half of 2023 has exceeded that in 2021. All of them provide strong evidence of people's attention to chatbots. Furthermore, we see from the figure that about one-half of them talk about bias, fairness, or ethics every year. It suggests that, as chatbots become more advanced, concerns about ethical issues also increase. Since the launch of ChatGPT, many papers have been published on this topic. Some analyze its bias and fairness in general [66, 149] while others are concerned with the same problem in specific applications. For example, ChatGPT may have political bias [154, 156], bias against conservatives [120], bias in healthcare and education [85, 158], etc. The power of LLMs has spawned many modern chatbots, and ChatGPT is only one of them. Although there are papers on bias in general chatbots, they only examine a narrow aspect, such as gender bias [63] and stereotypes [112]. They do not examine bias sources in chatbot applications in our society systematically. This is the void that we attempt to fill in this overview paper.

Although ChatGPT brings the ethical concerns of chatbots into the spotlight, bias in chatbots is actually not a new topic [161]. Most of the existing well-performing language models are ML-based models. ML algorithms face the bias problem in many aspects, such as data, user interaction, the algorithm itself, etc. [122, 179]. As a special ML system that interacts with humans directly, chatbots have a greater impact on ethical issues. To give an example, Microsoft released an AI-based chatbot called Tay via Twitter in 2016. It had the ability to learn from conversations with Twitter users. However, data obtained from Twitter users were seriously biased. Shortly after the chatbot was released, its speech turned from friendly, kind speech to discriminatory, offensive, and inflammatory speech in a short time. As a result, Microsoft had to shut down the chatbot urgently within a day after releasing it [133, 203]. Similar risks exist in recent chatbots. OpenAI CEO, Sam Altman, admitted that they were aware of ChatGPT's shortcomings in terms of bias in a Twitter thread in February 2023. Later, he added that technologies could "go quite wrong" and his "biggest fear" was that they would cause significant harm to the world. Some people with ulterior motives may take advantage of the flaws in chatbots and use them to harm society. To alleviate these problems, government regulation could be effective [32, 39, 109]. However, finding a balance between regulation and freedom of use is a problem that remains to be investigated, as over-regulation can hinder the development of innovation [216].

Ethical issues can be a barrier for companies to use large language models (LLMs) to interact with customers. In particular, the use of black-box models that lack transparency and interpretability to communicate with users is dangerous and unpredictable. A good chatbot can improve the user experience on the original basis, while a biased chatbot can cause a devastating blow to the user experience and cause serious damage. Recently, there are quite a few papers talking about the bias and fairness issues of ML systems, NLP algorithms, and ChatGPT applications. In contrast, there are fewer papers on bias and fairness in designing chatbot systems, which is the main focus of this overview paper. The main contributions of our work include the following.

- A comprehensive review of the history, technologies, and recent developments of chatbots.
- Identification of bias sources and potential harms in chatbot applications.
- Considerations in designing fair chatbot systems and future research topics.

The rest of the paper is organized as follows. The chatbot history, architectures, and categories are examined in Section 2. Possible bias sources, caused harms, and matigation methods in applications are discussed in Section 3. Considerations in designing a fair chatbot system are presented in Section 4. Future research directions are pointed out in Section 5. Concluding remarks are given in Section 6.

2 History, Architectures, and Development Categories of Chatbots

2.1 History of Chatbots

The history of chatbots is depicted in Figure 2. The concept of a chatbot was initiated in the Turing test in 1950. Various forms of chatbots have evolved over five decades in stages. Modern chatbots are built upon LLMs. The history of chatbots is briefly described below.

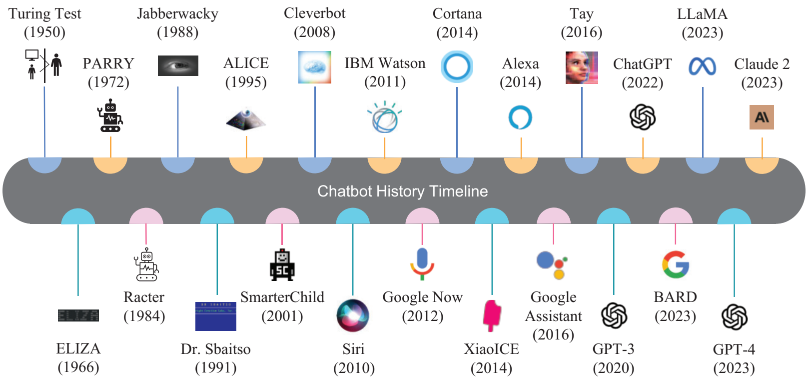

Figure 2: The history of chatbots.

1) Conceptual Stage. To answer the question "whether a machine can think?", Turing proposed the question-and-answer paradigm in 1950, which is known as the Turing test [185]. Simply speaking, one human participant would like to judge whether the other participant is a machine or a person through text questions and answers (rather than voice and/or appearance). If the human participant can hardly tell if the other participant is a machine or not, we may claim such a machine can think. The conversation idea between humans and robots through text was conceived in such a test. This is viewed as the origin of chatbots.

2) From 1960 to 1980. An early well-known chatbot, ELIZA, was developed by Weizenbaum at MIT to simulate a Rogerian psychotherapist in 1966 [200]. It found keywords from the input, reassembled user input and pre-prepared responses through certain rules to generate responses. ELIZA did not understand the meaning of the input during the process. It just conducted pattern matching and substitution. However, some of its responses made it difficult for people, who used the program for the first time at the time, to tell whether it was a machine or a human. Some people even developed an emotional attachment to it. The latter raised some ethical considerations [199]. Although ELIZA caused a sensation in the 1960s, it had many shortcomings. For example, its knowledge base was limited, and it could only answer questions in a certain narrow range. On the other hand, its appearance played an important role in inspiring follow-up research. Another famous early chatbot, PARRY, was developed to simulate a person with paranoid schizophrenia in 1972 [47]. PARRY interacted with ELIZA, who played the role of Rogerian's therapist. PARRY was considered more advanced than ELIZA because it had a better controlling structure and displayed some emotions [46, 210].

3) From 1980 to 2010. More explorations of chatbots were made from the 1980s to the 2000s. Racter was an AI program released in 1984. It generated prose in English, and its interactive version behaved like a chatbot. A learning AI project, named Jabberwacky, was conducted in 1988. It was designed to simulate natural human chatting in a fun way [167]. Different from earlier chatbots, Jabberwacky learned from chatting with people and stored keywords in previous conversations to grow its knowledge base [172]. Then, it used context matching with a dynamically growing database to choose appropriate responses [96]. Its new version, called Cleverbot, was released in 2008. Creative Labs designed a chatbot, named Dr. Sbaitso, for MS-DOS computers in 1991 and released it together with various sound cards in the 1990s [54]. Its interactive interface was a blue background with a white font. Although the interactive content was relatively simple, it used the speech synthesis technology and the sound card to realize text-to-speech (TTS) in the early stage. Inspired by ELIZA, Wallace developed ALICE (Artificial Language Internet Computer Entity) [193] in 1995. ALICE still used pattern matching rules but it was more capable since it had a much larger knowledge base. It used AIML (Artificial Intelligence Markup Language) to specify chat rules. Specifically, it used categories as basic knowledge units, where each category contains patterns and templates as user inputs and the corresponding machine responses, respectively [166]. ALICE gained significant recognition at the time. For example, it won the Loebner Prize three times in the 2000s [28]. However, it still failed the Turing test due to some limitations [170]. ActiveBuddy developed a chatbot, named SmarterChild, on the AIM platform in 2001. SmarterChild was one of the earlier chatbots that could help people with daily tasks through interaction, such as checking weather conditions, showtimes, stocks, etc. [6].

4) From 2010 to 2020. Watson was developed by IBM as a question-answering chatbot in 2011 [84]. It participated in the "Jeopardy" quiz show and won the championship twice. It was later used in the healthcare [43]. Microsoft developed a chatbot called XiaoICE [217] based on the emotional computing framework in 2014. It had both IQ and EQ modules and could flexibly answer user's questions. It was deployed in multiple countries and platforms. Another chatbot called Tay was released by Microsoft via Twitter in 2014. It learned from users but learned inappropriate remarks very fast, forcing Microsoft to shut it down shortly.

Furthermore, chatbots have been widely used in people's daily lives in the form of voice/search agents in instant messaging devices [87, 95]. Siri was released as an iOS app in February 2010 and integrated into iOS in 2011. It has been part of Apple's products since then. As a personal assistant, Siri can accept users' voice inputs and complete tasks such as making calls, reminding, looking for information, and translating [11]. Google released Google Now as a Google voice search app in 2012. It takes users' voices as input and returns with searched results. Microsoft launched Cortana for its Windows operating system in 2014. It responds to users' inputs using the Bing search engine.

Amazon launched Alexa, together with the Echo speaker, in 2014. Google launched Google Assistant and integrated it with Google Home speakers and Pixel smartphones in 2016. These voice assistants connect to the Internet and respond quickly. However, they face multilingual, privacy, and security challenges [25].

5) After 2020. The advancement of LLMs has impacted the development of chatbots greatly since 2020 [198, 215]. The transformer-based NLP technologies have made major breakthroughs in natural language understanding and generation [7, 37, 152]. LLM-based chatbots can provide rich responses using extensive training conducted on large pre-trained transformers. GPT-3 was released by OpenAI in 2020 and it laid the foundation of ChatGPT. ChatGPT was released in 2022 and gained more than 100 million users [205] shortly. Unlike previous chatbots, ChatGPT is an open-domain chatbot that can answer questions across a wide range of domains. LLMs have brought chatbots to a new level.

On the other hand, the popularity of ChatGPT has also led to a lot of controversy. As a large generative AI model, ChatGPT has a huge number of parameters. Its responses are difficult to predict and control, raising concerns about trustworthiness, toxicity, bias, etc. [219]. Its responses, which are highly similar to human beings, have aroused severe concerns. OpenAI released an even larger and more powerful LLM, called GPT-4, in 2023. The emergence of ChatGPT has impacts on the AI industry. In response to ChatGPT, Google launched Bard, a conversational generative artificial intelligence chatbot powered by LaMDA [182], in 2023. Meta announced its own LLM, called LLaMA [184]. Anthropic released Claude2.

2.2 Architectures of Chatbots

The architecture of a general chatbot is shown in Figure 3. It consists of five main modules: 1) user interface, 2) multimedia processor, 3) natural language processing, 4) dialogue management, and 5) knowledge base. The user interface module is responsible for input and output of the chatbot. It is the module that interacts with users directly. The input and output can be multi-modal. Multi-modal data is processed by the multimedia processor module. The NLP module is used to understand user's input language and generates the desired output language based on text answers. The dialogue management module is responsible for recording the current chat status and guiding the direction of conversations. It can access the database module and get answers for users. Because of the introduction of end-to-end LLMs, boundaries between various modules may not be as clear as those in the traditional chatbot design. Yet, there are still sub-modules that are responsible for the above-mentioned functions. The roles of these five modules are elaborated below.

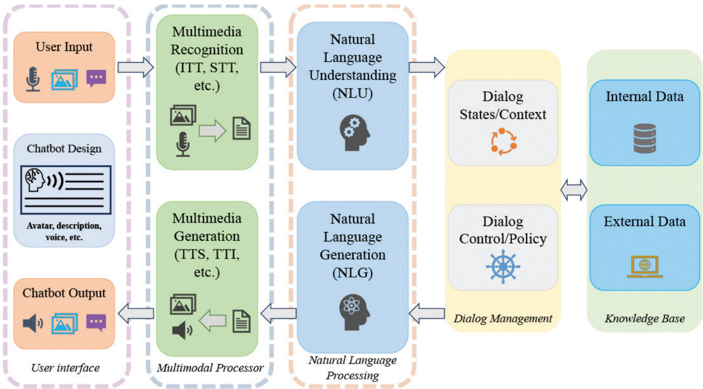

Figure 3: The architecture of a general chatbot.

User Interface. The user interface is a module that allows users to interact with the chatbot. It receives input from the user and provides output generated by the chatbot to the user. The system allows input and output of multiple modalities such as text, voice, or even pictures. Earlier chatbots used the text input. Voice assistants have been common in various devices nowadays. Recently, a variety of multimedia, such as images and videos, can also serve as input and output. To make chatbots more realistic, additional features, such as chatbot avatars, voices, emotions, etc., were added. These features can enhance user experience, making human-machine interaction smoother [9, 70]. For example, Microsoft's chatbot, XiaoIce, appears as an 18-year-old girl in a Japanese school uniform to users. She has a relatively complete resume and can do self-introduction. All these additions make the chatbot easier to be accepted as a virtual companion by users.

Multimodal Processor. In order to realize multimedia input and output (rather than text alone), chatbots need to convert data from different modalities to their corresponding text embeddings for further processing. For example, for speech input, the speech-to-text (STT) technique or the automatic speech recognition (ASR) technique [118, 132] can be used. For images, the image-to-text [113, 209] (ITT) technique has been developed. On the other hand, in response to users, chatbots need to convert text-embedded responses generated by the chatbot to various modalities, such as text-to-speech (TTS) or speech synthesis [135, 181], text-to-image (TTI) [72, 211], etc. In recent years, these techniques have been greatly improved because of the advancement of AI/ML [17].

Natural Language Processing. Natural Language Understanding (NLU) and Natural Language Generation (NLG) are two subtopics of Natural Language Processing (NLP). They are both key components in chatbot systems. NLU takes human text as input and converts it into a form that computers can understand and process. Two important tasks of NLU are intent recognition and entity recognition [91]. Intent recognition refers to understanding user's intention and observing user's emotion. It serves as the basis for chatbots to generate reasonable answers. Irrelevant answers degrade user experience significantly. Entity recognition refers to the extraction of entities that exist in real life in user input sentences, such as objects, people, cities, etc. These entities help the chatbot make logical reasoning and find answers in the knowledge base. NLG is another important component, which is responsible for generating human language fluently and naturally [75]. Using NLG, computers generate emotional responses using human natural language, thereby enhancing user experience and trust.

Dialog Management. Dialogue management (DM) is used to decide the communication strategies [33]. It needs to remember the current dialogue state and control the content of the next dialogue. After acquiring user intent and input entities, the DM module analyzes them, records the current dialogue state, and then decides the direction of the dialogue according to the contextual dialogue states. For example, if an entity needed to answer a question is missing, the chatbot will ask the user to provide more information. DM also needs to record some information (e.g., user preferences, dialogue background, etc.) and use certain logical reasoning to give appropriate responses. Although DM and NLU are independent modules, they affect each other's performance in many ways [82].

Knowledge Base. After clarifying user's intention and obtaining the necessary information, DM accesses the knowledge base to obtain the desired answer. The source of the data in the knowledge base can be the internal knowledge stored in the chatbot or the external knowledge available through the Internet. Data is stored in a graph-structured format in the knowledge base, where nodes are the entities and edges are the relations. The design of knowledge bases facilitates fast, accurate, and reliable reasoning to help DM locate the correct answers efficiently. Several graph machine learning algorithms, such as multi-hop reasoning [2, 76, 206] and graph neural networks [98, 188], can be adopted and improve the chatbot performance even further.

2.3 Categories of Chatbots Based on Development Methodology

There are three main categories of approaches to develop chatbots: rule-based, retrieval-based, and LLM-based [88, 183]. Their main differences are summarized in Table 1, which will be elaborated below.

Rule-based. Rule-based chatbots look for keywords in user input and respond using predefined rules. They are adopted by simple question-and-answer systems. Early chatbots were mainly rule-based, such as ELIZA and AIML-based chatbots [160]. In developing rule-based chatbots, the design team needs to define rules manually, which is a tedious job. Since the content that humans can define is limited, input queries that can be effectively replied

Table 1: Three chatbot categories based on their development methodology.

Chatbots	NLU	NLG	Bias
Rule-Based	Pattern Matching	Predefined	Low
Retrieval-Based	ML Algorithm	Predefined	Low
LLM-Based	ML Algorithm	ML Algorithm	High

to are limited. In the face of situations where keywords cannot be matched, rule-based chatbots can only change the topic using predefined sentences. The decision-making process of rule-based chatbots is clear, and their responses are controllable. Rule-based chatbots have a poor understanding of context and language. Their answers lack novelty and could be highly repetitive. Since keyword matching and response content are all set in advance, the bias primarily comes from the development team. The bias level is relatively low.

Retrieval-based. Like rule-based chatbots, retrieval-based chatbots only give predefined answers so their answers could be repetitive. On the other hand, such chatbots have learning capabilities. That is, they use machine learning methods to train part of the question understanding system. Thus, they can choose more appropriate answers from existing ones. Besides biases from the development team, retrieval-based chatbots are subject to biases arising from machine learning. However, since their responses are predefined, the bias problem is more manageable by humans.

LLM-based. Unlike the previous two types of chatbots, LLM-based chatbots can generate new responses using large language models. They use ML algorithms to understand user input and generative AI (GAI) algorithms [197] to generate responses with a certain degree of randomness. For example, ChatGPT is an open-domain chatbot that can answer users' questions in different fields. They face several challenges at the same time. First, training such chatbots requires a lot of data and computing resources, which is costly. Second, the mainstream LLM-based chatbots use unpredictable black-box models. They are uninterpretable and without a logical reasoning process. Consequently, their responses are unpredictable and difficult to control. They often contain inappropriate or fake content with biased and offensive language, etc.

3 Bias Issues in Chatbots

The deployed chatbots contain biases from various sources [18]. We categorize them into three types for ease of analysis. As depicted in Figure 4, they can arise from: 1) chatbot design, 2) user interactions, and 3) social deployment. First, a chatbot development team is made up of people from different educational and cultural backgrounds. Their personal biases will have an impact on the designed

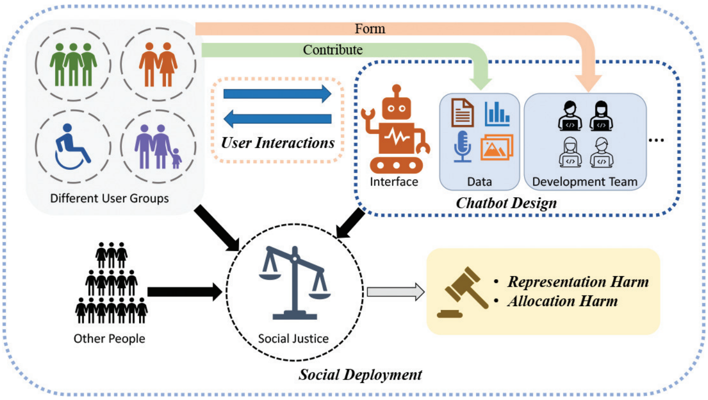

Figure 4: Three bias sources in a chatbot system.

chatbot system. A chatbot system is composed of an external user interface and several internal modules. The user interface design is directly influenced by the development team. The internal modules rely on training data and ML algorithms. Different data source groups contribute to data acquisition and annotation, which can be biased as well. The final data used for training needs to be screened by the development team, leading to another bias source. Second, after a chatbot is deployed, it interacts with users and biases can be enhanced in the interaction process. The bias can even affect user's view and value. In addition, users may become part of the development team and contribute to data annotation in the future, which makes bias generation a vicious circle. Third, biases may come from the environment where chatbots are deployed. For example, people's attitudes toward chatbots and the way chatbots are used can lead to biases. A biased chatbot system and people affected by the bias may result in representation and allocation harms to social justice. These topics are the main focus of this paper. They will be detailed below.

3.1 Biases from Chatbot Design

Figure 5 gives an overview of biases in designing chatbot systems. A chatbot is composed of a user interface module and several internal components. Each of them can have several bias sources. The development team may pass its biases to each component. On the other hand, they can utilize some toolkits to mitigate biases.

3.1.1 Biases in Development Team

People are biased [34], and developers are no exception. Individual biases may be influenced by personal experience, family upbringing, culture, education, etc.

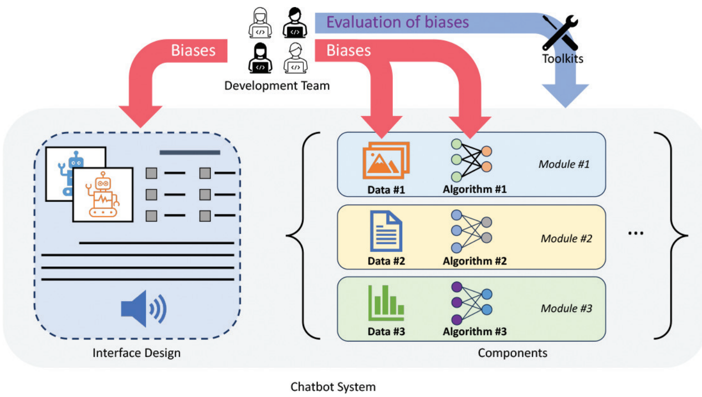

Figure 5: Biases in the design of chatbot systems.

Developers' cognitive biases can affect developed software [127]. Our brains tend to simplify the world for decision efficiency, which results in cognitive biases [105]. Being subject to cognitive biases, developers may ignore certain factors and/or situations in designing rules or selecting training data and/or input features. Another example is that an all-male development team often sets the chatbot gender to a female. Some of the biases are explicit, while others could be unconscious and implicit. For user groups with similar biases as developers, there is an attraction effect between them and the development team [207]. They may have a better user experience with the chatbot. On the other hand, implicit or unconscious biases may worsen the user experience of other groups.

One way to mitigate such biases is to increase the diversity of the development team [63, 123]. Although it would be ideal to include subgroups among users to reduce the marginalization of minorities, this is however too difficult to implement. In practice, the development team may hire people of diversified backgrounds, with different ways of thinking and perspectives, and with different expertise to control biases. Besides, the development team can use some toolkits to evaluate biases of the system and take steps to mitigate them. For example, FairPy [192] is a toolkit for quantitative bias evaluation in pre-trained LLMs like BERT. It takes different types of biases into account, such as race, gender, age, etc. AI Fairness 360 [19] is another toolkit aiming at transitioning fairness research to industrial settings and providing a general framework in fairness algorithm sharing and evaluation. It helps developers detect and mitigate biases in ML models. As another toolkit, Aequitas [157] takes application scenarios and non-technical people into account. It sets up several bias and fairness metrics associated with multiple population subgroups in the ML workflow. As a scalable toolkit, LiFT [189] can incorporate the bias measure and mitigation mechanism in ML systems operating on distributed datasets.

3.1.2 Biases in Interface Design

Unlike general ML systems, chatbot systems have unique interfaces. To make chatbots more human-like and improve user experience, companies add some attributes to chatbots, such as names, resumes, descriptions, avatars, and voices. Most users do not understand the architecture of chatbots and can only interact with chatbots through these attributes. Then, the interface design has an impact on user perception. However, a chatbot's attributes can contain stereotypes and biases, with the gender bias being the most prominent [48].

A UNESCO report [201] pointed out that the vast majority of voice assistant chatbots (e.g., Siri, Alexa, etc.) are designed to be female. They have feminine names with female voices and appearances as the default. A study [63] examined 1,375 chatbots on the chatbots.org website and found that most chatbots were designed as a female. This is especially true in the sector of customer service and sales. In real life, many workers in the service industry are women [150], and people often have stereotypes of women as helpful, gentle, accommodating, and nurturing. When chatbots appear in people's lives as service providers, developers set them as women by default to increase affinity.

On the other hand, some robots, such as those designed for stereotypically male tasks that need to show strength and leadership, use male personas [49]. Due to people's gender stereotypes, names, voices, body shapes, and facial cues of robots may affect user's impression and trust in robots [20, 21, 62, 131]. They may have an impact on customer satisfaction [165]. Such a setting may be in company's business interest, since users prefer chatbots that present stereotypes of specific roles [119]. Another reason for chatbot feminization could be the low proportion of women in the chatbot development team [69]. Apparently, designing a chatbot to be a female is more appealing to an almost all-male development team.

However, the design, which is beneficial to business, could be harmful to our society. The ubiquity of chatbots designed according to stereotypes will reinforce people's stereotypes. For example, in the case of voice assistants, the feminization of voice assistants contributes to stereotypes about women. As chatbots, they will try their best to meet the needs of users. However, female chatbots are more likely to be targets of sexual harassment and abuse [29, 204]. When faced with inappropriate requests such as sexual harassment and bullying, most of them choose to avoid or pretend not to understand [51]. These responses appear to set an example for women to accept abuses and teach them how to respond to unjustified demands. This kind of response turns the sadistic behavior into an acceptable behavior and reinforces the stereotype that women are accommodating and submissive. The consequence clearly brings harm to our society. Nowadays, some companies are aware of this issue and have taken actions to control the gender bias in chatbots, such as providing male voices instead of the default female voice and making chatbots appear tough in the face of inappropriate language. However, the interface design is still a source of biases in chatbot systems. It can contain many forms of biases, of which the gender bias is an obvious one.

3.1.3 Biases in Internal Components

A chatbot system consists of many internal components such as multimodal processors, natural language processing, dialogue management, etc. Each is responsible for a specific task as shown in Figure 3. The design of each component can lead to biases. For rule-based modules, biases are mainly from limited rules and predefined responses. They are less and easier to control. For ML-based components, the use of ML algorithms makes the system more capable yet biased. ML algorithms are used in a wide range of fields [159], including important and sensitive areas such as healthcare [124] and recruitment [102], nowadays. Their bias problem has received more attention. There are many papers on the bias and fairness issues in AI and ML. Some analyze possible bias sources in general AI systems [134, 153, 174], and attribute them to data sources and AI/ML algorithms. Others discuss specific types of bias in AI systems, such as the racial bias [106, 137] and the gender bias [59, 129]. There are also papers on specific application fields such as healthcare [45, 141] and education [15, 99]. Generally speaking, the bias in the ML-based modules mainly comes from data and algorithms, e.g., data collection and labeling, feature selection, the way data are used in ML algorithms, etc. [122]. Biases in one component can have an impact on the performance of the other, and the contribution of each component to the overall biases in the system is often difficult to determine. Combinations of individual components that meet fairness metrics can also exhibit biases. To control the biases of the whole system, it is important to mitigate the bias in each component and evaluate the biases of the whole system.

Biases in Data. The bias may come from the data source or from people's collection and labeling of data. The data source is affected by human biases such as reporting bias, selection bias, etc. [77, 202]. It is recorded for various reasons. It may neither reflect the actual distribution nor have a proper balance among subgroups. Next, training data are selected from the data source and annotated. In the selection process, there is sampling bias, in-group bias, measurement bias, etc. [53, 68]. Thus, the distribution of training data may not be the same as that in the application scenario. Furthermore, the data labels can be affected by the world views of annotators, resulting in experimenter's bias and confirmation bias [100].

Biases in ML Models/Algorithms. Features considered by ML models often lead to biases [104]. Even if sensitive features are not directly used as input, some features that are highly correlated with sensitive features will allow models to learn bias. During training, biases in the training data are amplified. In a general ML system, the aggregation bias [178] may occur under the influence of broad categories of data groups, and models may draw wrong conclusions about individuals due to group trends. In addition, the evaluation of models can also lead to bias. Sometimes, the evaluation criteria do not accurately reflect the desired goal. In a chatbot system, almost every module can be implemented using ML algorithms to improve performance, so they may all contain certain biases.

Biases in Multimodal Processors. As AI systems become more advanced, people are no longer satisfied with the input and output of a single modality, and multimodal processors are used to handle the multimodal communications between humans and robots. The combination of multiple modalities will often increase model biases and compromise fairness [27]. Biases can be hidden in both algorithms and training data for each modality. For example, the automatic speech recognition (ASR) technique, which enables chatbots to recognize and interpret human speech, is an essential component of voice assistants. However, ASR systems can have biases such as gender, age, and regional accent biases [64]. They may come from the composition of the corpus, the mismatch between the pronunciation and speech rate of users and the training data, or biased transcriptions, etc. Another example is image captioning. The ML model converts images to text, which can be affected by biases in the image context. For example, men are associated with the snowboard while women are associated with the kitchen in training images [83]. In chatbot systems, multimodal processors are usually directly connected to the interface. When converting multimodal user input into text, their biases may distort user input or omit important information, which affects subsequent dialogue understanding and answering. On the other hand, when converting generated text to multimodal output, the bias may lead to inappropriate output content and affect user experience.

Biases in NLP Models/Algorithms. NLP is a branch of AI that aims to equip computers with the ability to understand and generate human languages. There are quite a few recent papers on the bias and fairness issues in NLP [24, 40, 58, 60, 74, 86]. They show that NLP models can contain biases and there are methods to mitigate them. For example, word embedding is an important technique that represents words in vector form to facilitate computer understanding of human language. However, word embedding models often contain human-like biases [1], such as associating men with computer programmers while associating women with homemakers [26]. The use of word embedding in NLP downstream tasks may amplify certain biases. The cosine similarity-based Word Embedding Association Test (WEAT) [36] and its variants [55, 79, 111] have been proposed to measure and mitigate these biases. They are applied to models such as Word2Vec [125] and GLOVE [145].

Besides word embedding, there are other biases in NLU and NLG tasks. For the NLU task, biases in coreference resolution have received much attention [155]. Methods like debiased word embedding and data augmentation can mitigate these biases effectively without affecting performance much [142, 213]. For NLG tasks, biases can come from deploying systems and decoding techniques [168]. Methods for measuring and controlling such biases have been proposed as well [57, 143, 169].

With the advancement of NLP technology, large-scale language models are trained on a wide range of corpora to master general human languages. Although these pre-trained models can be used as starting points for downstream NLP tasks to improve efficiency, they can be biased [162]. For example, as a large-scale natural dataset in English, StereoSet [130] evaluates biases in pre-trained models and demonstrates that most mainstream pre-trained models exhibit strong stereotypical biases. These biases are propagated to downstream tasks, affecting the performance of downstream models.

3.2 Biases from User Interactions

In the chatbot development phase, biases mainly originate from developers, training data, and algorithms. When the service is launched, the chatbot interacts with users. It gets prompts and feedback from users. It learns from interactions, which makes it more capable but introduces bias. This is a significant difference between chatbot systems and other traditional ML systems. The interactions between a chatbot and users are illustrated in Figure 6. Users first give the chatbot a prompt to start a conversation. The chatbot will return with a response. Users can grade the response, and ask for regeneration or give a new prompt. In this way, users and the chatbot can exchange information with their biases being propagated mutually. This topic is elaborated below.

Biases in User Prompts. Unlike some ML systems where users accept predictions unilaterally, chatbots respond based on prompts from the user. Users' prompts can be in multiple languages. Automatic determination of the language is an important step for further processing. Language identification (LID) is usually used to detect the input language type [121]. An ideal LID tool should be unbiased to any language in terms of inference time, response content [8], etc. As pointed out in [16], dialects can cause biases in NLP tasks. For prompts that contain informal language, such as dialects, chatbots may misunderstand and respond with biases. Also, keywords in prompts are obvious. For chatbots that learn stereotypes from massive data, certain keywords may trigger stereotypes in the model and make the model generate stereotyped results. For example, a prompt containing the keyword "Muslim" might yield results related to "Terrorist" [3]. Sometimes, although there are no obvious keywords in the prompt, specific chat topics can lead to bias [171]. For

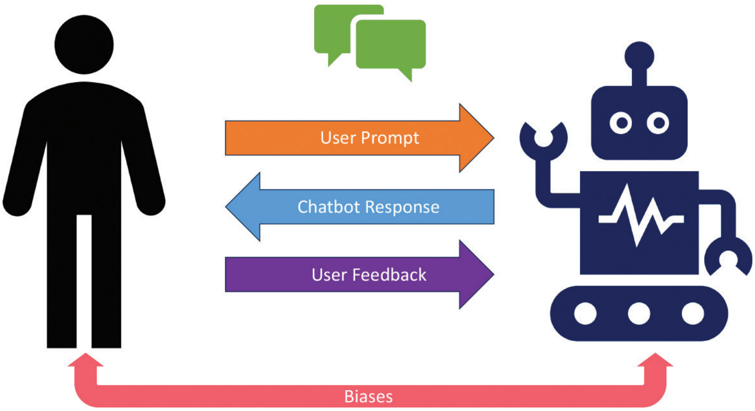

Figure 6: The bias coming from the interaction of users and a chatbot.

example, women are more likely to appear in the topic of "family" in GPT-3 [116]. Also, input of similar meaning in different prompts may lead to different results. A prompt could be biased, and a chatbot that cannot perceive the bias may respond to biases given by the prompter. Even if the prompt is not biased, the chatbot may show a certain tendency. To get fairer results, chatbots should be able to recognize biases in user prompts. Users may need to think about chatbots' derivation process instead of accepting the responses blindly.

Biases in User Conversation. Learning from conversations with users is an important way for some chatbots to expand their knowledge base and become more powerful. On the positive side, this can enhance their understanding of human language and allow them to generate more accurate responses. However, the user group is large and complex, and user prompts can contain biases and/or stereotypes. Some users may even deliberately guide chatbots to express biased speech. If the chatbot learns in a wrong way, it could be harmful. Microsoft's chatbot Tay was an example. The rapid change of its conversational style demonstrated the destructiveness of continuous learning from biased user conversations. Thus, it is important to check the responses generated by the chatbot with a systematic approach at a certain frequency after its initial deployment.

Biases in User Feedback. Many systems allow users to give some feedback on the response, such as asking to regenerate the response or rating the response after a chatbot responds. This can improve the user experience. The feedback can help chatbots understand what kind of results are more accurate and more in line with users' expectations to enable the machine to generate more satisfying or personalized responses for users in the future. When chatbots generate biased, fabricated, or incorrect responses, negative feedback from users can help them correct mistakes. Without feedback, chatbots may not be able to detect and correct their own biases timely, and these biases may be further amplified in subsequent learning. However, users have various biases and people favor responses that cater to their biases [119]. As a result, they tend to give higher scores to responses that contain biases. When chatbots use biased user feedback to learn and improve responses, future responses will have increasingly severe biases. Thus, how to use user feedback to alleviate biases in chatbot systems is an important topic.

Biases in Chatbot Responses. Biases are passed on to users in chatbots' responses. Generated responses of some chatbots are always confident and smooth, regardless of whether they are correct and/or may contain inappropriate information [180]. This makes it difficult for users to distinguish authentic information from falsified information. Young children, in particular, lack the ability to recognize biases in chatbot responses and to judge truth from fiction. It thus poses a challenge to chatbot's educational applications [94, 103]. One solution is to ask the chatbot to provide references and the reasoning process along with their responses. The additional information helps users judge whether the content is reasonable and correct. However, some modern chatbots, such as ChatGPT cannot provide proper references [78]. For example, when users ask ChatGPT to give reference links, most of them are actually irrelevant [52]. The opacity makes it difficult for users to judge the accuracy of received responses.

Presentation and Ranking Biases. Chatbots play a role similar to search engines that may have biases in delivering information to users [13]. Unlike traditional search engines, they extract and summarize the information on the Internet (rather than listing raw URLs). Since there is too much information on the Internet, it is difficult for chatbots to present all of them. Which information to present and whether the presented information is balanced are determined by the algorithms. The information not presented cannot be received by users. All of these lead to presentation bias. The presented information by chatbots may be ranked or with a certain focus, causing ranking bias.

Vicious Bias Circle. When people have long-term conversations with biased chatbots, the passed biases can affect their worldviews. This is especially severe for children. The biased worldviews will affect data collection and annotation, model training, and chatbot development. In this way, biases will become more serious, forming a vicious circle as shown in Figure 7.

3.3 Biases from Social Deployment

Technologists try to mitigate biases in the product design to make the chatbot fairer. However, failure to understand the interaction between the chatbot and its social environment may result in the contextual mismatch [164]. The

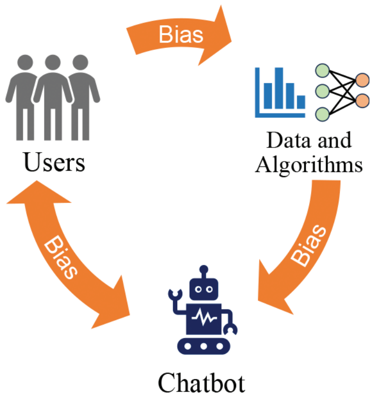

Figure 7: Illustration of the vicious bias circle in chatbots.

chatbot is deployed in a community as shown in Figure 4. There are multiple user groups interacting with the chatbot. Some of them may never use the chatbot before. Different chatbots have different social contexts, and their fairness consideration cannot be separated from the social context. Everybody in the community has the potential to influence the chatbot and, in turn, to be influenced by people who interact with the chatbot. Chatbots that meet the fairness criterion before deployment may become biased in a particular community later since biases may stem from people's attitudes to chatbots, different user group compositions, different chatbot usages, etc. Thus, it is essential to consider the chatbot deployment environment to mitigate the bias. We will focus on biases arising from people's attitudes, application background, and solution selection below.

Biases in People's Attitudes. When a new technology is applied in a social environment, human perception of the technology may lead to biases that have a great impact on human-machine cooperation. Even if a technology meets fairness metrics, whether it will be biased in practice depends on how people use it. The same is true for chatbots. User's attitudes toward a chatbot determine how they use it and the impact it can have. Since chatbots reason differently from humans and make different mistakes, they would perform better when humans and chatbots work together. On the other hand, a chatbot's speech may unduly influence human behavior, and the way people use chatbots may go beyond the expectations of their designers.

Some people may unconsciously believe in chatbots for various reasons, such as not being confident enough about themselves, over-believing in the answers of chatbots, fear of taking responsibility [194], etc. All of them lead to the automation bias, which allows chatbots to propagate wrong knowledge or fake content more easily. Some of today's chatbots speak in a very confident tone, regardless of whether the content is correct or not, which exacerbates this phenomenon.

The automation bias is common in AI systems. For example, results of the Allegheny Family Screening Tool (AFST), a predictive model for child abuse, may cause staff members to question their own judgment [61]. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which predicts the risk of recidivism, is another example. Its designers did not intend for the model to determine a person's prison time. Yet it was used by judges in sentencing [10]. People tend to rely too much on AI systems because they may subconsciously think that robots without emotions would give unbiased conclusions. They doubt themselves and follow chatbot's opinions when they have different viewpoints. Similar issues arise in chatbot systems that are used to predict or provide decision-making advice.

The chatbot output is in human language. It can be used to write articles. If an author relies too much on chatbots, the resulting article may contain fake or plagiarized content. The opaqueness of sources and the flamboyance of the article make it difficult for authors to spot the errors. Once caught, the author may attribute the error to the chatbot. However, chatbots should not be held accountable for the mistake. It is important to have an appropriate accountability mechanism in place [187].

Besides, people may have an aversion to chatbots. This happens for a number of reasons. For example, some might be happy working with a chatbot at first. However, after the chatbot made severe mistakes, they did not want to continue working with any chatbot. Others may be influenced by media propaganda to form stereotypes about chatbots. Furthermore, some do not trust third-party chatbot companies and cannot accept their domain being influenced by them. Others worry that chatbots will take their jobs [81] and try to exclude them as much as possible. Without the aid of a fair chatbot, people may struggle to detect their own implicit biases. Whether believing in the chatbot system too much or resisting it too much is not helpful to building a cooperative relation between humans and chatbots. Understanding and analyzing how people perceive chatbots in social environments are needed to control potential biases.

Biases in Application Domain. Different chatbots are designed for various applications [90]. Open-domain chatbots can talk to people without being limited by topics and domains. Domain-specific chatbots master the knowledge in specific domains, and they are designed for specific tasks. Their social environments and served user groups can be quite different. Three key factors of deploying chatbots in a specific application are shown in Figure 8. They are social needs, design goals, and actual effects. Before designing a chatbot, the development team should understand social needs and establish appropriate design goals accordingly. Then, the chatbot will be designed based on the design goals. The final effects of the chatbot product should meet design

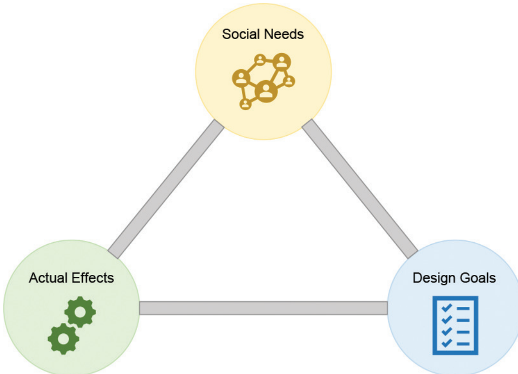

Figure 8: Three key factors of deploying chatbots in a specific application domain.

goals and social needs. The three factors should be consistent to minimize the biases in a specific social environment. Design goals play an important role as a mediator between actual effects and social needs.

To deploy a chatbot in a social group, understanding social needs is the first priority. The development team should set up the design goals and choose proper fairness metrics according to the needs so as to design a chatbot that can be well integrated into the society. However, in reality, programmers tend to emphasize portability. Some models may be shared under different social needs, such as many pre-trained LLMs in NLP. Sometimes, models may be designed for a specific user group, and their design goals are based on some assumptions about the social context. When transplanted to the other user group, these assumptions may not hold and there is inconsistency between design goals and social needs. Then, biases could appear. For example, a chatbot is designed for users in a country that has specific training data and design goals. If the same chatbot is deployed in a more conservative country, its responses may be considered offensive and not in line with user expectations. Clearly, the conflict between social needs and design goals leads to biases.

Inconsistency between design goals and actual effects can lead to the bias. The development team needs to model and implement a chatbot according to design goals, including data collection, feature selection, rule-making, social equity modeling, etc. Improper modeling may result in poor models. For example, a development team wants to design a companion chatbot that understands and responds to user emotions. The design may allow the chatbot to understand users' emotions through real-time facial emotion recognition. However, facial expressions may have different emotional implications in various cultures. The use of facial expressions alone to judge emotions is problematic in practical applications [42].

Table 2: Comparison of the allocation and the representation harms.

Allocation Harm	Representation Harm
Short-term	Long-term
Limited range	Wide range
Evident	Elusive
Readily quantifiable	Challenging to quantify

Biases in Solution Selection. When people get used to a solution, it is often difficult to think of new ways to solve the problem. For a new problem, development teams will naturally give priority to the solutions that they are more comfortable with. This bias may come from propaganda in the society or people's experience. For example, the recent popularity of ChatGPT has motivated people to apply it to various fields. LLM-based chatbots have become mainstream solutions to many problems [12, 80, 195]. However, they may not be optimal for some tasks. For example, traditional chatbots with predefined output results may be more economical for customer support. They save resources and have a lower risk in the bias. While results generated by generic models such as ChatGPT are diverse, they may not be accurate and correct. For chatbots in the medical field, although LLMs such as ChatGPT are applicable, their opaque reasoning process can be challenged. Besides, there is a risk in the leakage of personal sensitive data during the dialogue process between humans and chatbots. Traditional methods could be more robust and privacy-preserving. While advanced methods marginalize some user groups, traditional methods can be tailored to them and offer a better user experience.

3.4 Harms from Negative Biases

Biases may not always be bad, but negative biases can result in serious harm to our society. Harms caused by biases can be divided into two types: the allocation harm and the representation harm [50]. The allocation harm occurs when the machine is used to make decisions and allocate resources, which significantly benefit some groups against others. The representation harm occurs when the machine reinforces associations or stereotypes between certain representative traits and some people groups. The differences between the two are summarized in Table 2.

Allocation Harm. The allocation harm often has a specific application scenario and happens quickly. It happens when decisions made by people are influenced by biased machines. As chatbots become smarter, people may rely on them to analyze problems. If chatbot's analysis is biased and influences human decisions, the allocation harm arises. For example, existing research suggests that using machines to guide health decision-making may result in the allocation harm to African-Americans [138, 148]. With the popularity of LLM-based chatbots, we see a trend to use chatbots for automatic diagnosis. When a biased chatbot is used for diagnosis, different age or ethnic subgroups with the same symptom may have different diagnostic recommendations. This affects people's judgment on the condition and subsequent medical decision-making, leading to the unfair distribution of medical resources. The allocation harm leads to the short-term unfair distribution of social resources and affects a limited range of people. It is easier to detect and quantify.

Representation Harm. The human society has many stereotypes that associate representative characteristics with specific groups. In the context of chatbot systems, users communicate frequently with chatbots. If some characteristics are always associated with specific groups in the conversation with the chatbot, users' worldviews can be subtly changed and stereotypes will be exacerbated. The wider the user base, the more far-reaching this effect will be. The representation harm also occurs in chatbots when affected users participate in developing chatbots, such as data recording and annotation, model design, and system evaluation. The representation harm takes a longer time to develop and lasts for a longer time. It is elusive and relatively difficult to detect, change, quantify, and track. Although the representation harm is less obvious, it cannot be underestimated. It may change people's perception of the world and the future direction of our society.

3.5 Bias Mitigation in Chatbots

As people pay more attention to biases in AI systems, various methods to mitigate biases have been proposed. They are generally categorized into pre-processing, in-processing, and post-processing of ML algorithms. Similar methods can also be effectively applied to address biases in chatbot systems. For chatbot systems, biases can be mitigated in three stages: 1) preparation, 2) development, and 3) optimization. In this section, we will explain how biases can be mitigated in each stage of chatbot systems.

Preparation. To design a chatbot system for a specific application, background research in the preparation stage is important. Understanding the context of the application helps identify the social needs, set reasonable design goals, and assemble a balanced development team. It mitigates the biases caused by contextual mismatch. Preprocessing is important in the preparation stage to mitigate underlying biases in the training data by aligning the statistical distribution of the datasets with the real-world application scenarios. For example, several data preprocessing techniques, such as oversampling, undersampling, data augmentation, etc., can be used to mitigate the bias in the training data [147]. Removing sensitive features and relabeling some samples in the dataset before training also help mitigate biases [115].

Development. During chatbot development, developers need to design a proper user interface and algorithms for different internal components. For the interface, suggestions and feedback from experts and volunteers in different user groups are important for designing and evaluating. To mitigate biases from internal components, adopting models that have better performance on specific fairness metrics is beneficial. In addition, having an interpretable and transparent decision-making process is crucial to bias alleviation. After choosing certain models, techniques are available to further mitigate biases. For example, to mitigate the gender bias in NLP tasks, learning gender-neutral word embeddings, using constrained conditional models, and utilizing adversarial learning are helpful [177]. More generally, adopting fairness-aware classifiers [208], adding regularization and constraints, and ensemble different models are useful in mitigating biases in ML algorithms. In the development phase, the fairness toolkits mentioned in the previous section are useful for developers to evaluate their algorithms and make adjustments.

Optimization. As an AI system that primarily interacts with humans, chatbots require further optimization and maintenance after deployment [128]. When communicating with users, biases that were not considered may occur [173]. Developers need to optimize the system and mitigate such biases to prevent further harm. For example, they can use a rule-based model or train a new module to detect biased prompts from users. Once the model finds a biased prompt, the chatbot can correct the prompt or deny the request. When learning from human interactions, human supervision or bias detection models can be introduced to filter out biased information so the chatbots will not learn from the misinformation. Besides, after a system is deployed, it is important for developers to explain the proper usage of the chatbot and inform users of potential risks and the scope and capability of the chatbot. Developers may also check the misuse problems to avoid biases regularly.

4 Fairness in Chatbot Applications

With the advancement of AI and the occurrence of many unfair cases caused by AI applications with bias and discrimination [65], people have paid more attention to the fairness of AI in real-world applications [38]. As a specific AI system that interacts with humans, fairness issues arise in chatbot applications [114, 212]. A fair system means one without negative bias and discrimination. An unfair chatbot system may produce biased or discriminatory responses against certain individual users or user groups, leading to the spread of biases and stereotypes and causing harm to the society. While the bias can sometimes be unintentional and arise from a variety of factors, fairness is an intentional goal that people strive to achieve. To judge whether a system is fair, we need to define fairness first. It is, however, difficult to give a universal definition of fairness. Fairness is a complex concept that depends on the application context. Different groups and individuals see fairness differently. Multiple definitions of fairness are discussed in [38, 122, 126, 191]. Each of these definitions represents unique perspectives on the applications and the interests of different groups.

Most fairness definitions can be roughly divided into four categories: group fairness, individual fairness, causal fairness, and counterfactual fairness. They are elaborated below.

- Group fairness, such as equal opportunity and demographic parity, focuses on differences between the chatbot's responses to different groups given similar prompts. It aims to eliminate discrimination against certain groups. For example, when a chatbot is asked to create jokes related to a specific racial group using stereotypes, it should reject such requests consistently across all the groups. Group fairness cannot be achieved if requests regarding certain racial groups are denied while requests regarding other racial groups are accommodated.

- Individual fairness, such as fairness through awareness, emphasizes differences between responses received by individuals of similar backgrounds. For example, if a chatbot is asked to provide rehabilitation recommendations, patients with similar conditions should receive similar recommendations rather than completely different recommendations based on demographics such as gender or age. If a male patient is advised to exercise more and eat more protein, while a female patient is advised to rest more, the chatbot does not meet the goals of individual fairness.

- Causal fairness evaluates the fairness of a system from the perspective of changing a specific characteristic and observing biases in the response [73]. It aims to mitigate the impact of specific attributes on decisions and avoid the system perpetuating historical biases and inequalities. For example, if a chatbot is asked to recommend products, it should give balanced suggestions to all users based on individual preferences. If the chatbot primarily recommends cosmetics to women and electronics to men only based on historical data, it may propagate historical gender-based inequalities.

- Counterfactual fairness [108] evaluates fairness by considering hypothetical scenarios where sensitive attributes are different while other attributes are the same. For example, if someone asks for chatbot assistance with a hiring decision. The person could then create a counterfactual scenario by switching the gender of the applicant to see whether the advice given by the chatbot is consistent. If they are inconsistent, counterfactual fairness is not met.

Each fairness definition is reasonable. However, satisfying all of them at the same time is challenging [101]. To design a fair chatbot, it is crucial to clarify its application context, including its purpose, who will use it, how it will be used, its difference from traditional methods, people's attitudes, and possible loopholes. In the development process, designers should consider additional questions, e.g., what sensitive features may be implied in the prompt words, which fairness definitions should be selected and their priority, whether current fairness considerations will change over time, what causes the differences between groups or individuals and whether they are reasonable, and so on. In addition, designers should think about the consequences of false positives and false negatives. When a chatbot makes mistakes, which of the two will have a more serious outcome? For example, for a chatbot used for children's education, the impact of not blocking inappropriate content by mistake is much worse than blocking irrelevant content by mistake.

5 Future Research Directions

5.1 Open-domain Versus Domain-specific Chatbot

Open-domain chatbots, such as ChatGPT, have demonstrated their power recently. They can handle prompts in multiple domains of complex social contexts and from a wide range of user groups. Generally speaking, it is difficult to mitigate biases and develop fair open-domain chatbots. Domain-specific chatbots are different. They have specific user groups and preset application scenarios. Thus, it is easier to consider possible biases, choose the appropriate fairness metrics, and implement a relatively fair system. Besides, open-domain chatbots with LLMs usually require a huge amount of training data and computing resources. While domain-specific chatbots have limited knowledge, they demand much less training data and computing resources. They are easier to control. In some fields where accuracy or user privacy is important, such as healthcare, domain-specific chatbots are more likely to obtain accurate responses than open-domain chatbots under the same amount of resources. It is also easier to protect users' privacy in domain-specific chatbots with a smaller model size that runs locally without uploading data to the public server.

5.2 Bias Control in Multi-Modal Chatbots

Chatbots with multi-modal input/output will become the main trend in the future. Such chatbots not only need to take care of NLP and dialogue management, but they also need multiple models for modality conversion and integration. To realize a fair chatbot system, modality conversion and integration models have to satisfy their respective fairness metrics. After putting them together in the system, they may affect each other and the overall output may be biased. It is important but challenging to mitigate the bias of the whole system.

5.3 Green and Interpretable Chatbots

LLM-based chatbots become popular recently. Since LLMs need a huge amount of computing resources to train, they are not environment friendly. The chatbots are built upon large pre-trained models with fine-tuning. They contain biases from multiple sources. They are difficult to mitigate since developers treat the whole system as a black box. To detect and mitigate biases, one can only rely on input prompts and output responses. As a result, the bias detection job is labor intensive. On the one hand, traditional chatbots contain less bias. On the other hand, they are inferior to LLM-based chatbots in generating fluent human natural language with rich and diversified content. It is appealing to design a logically transparent yet content-rich chatbot.

One possible direction is to leverage the tool of knowledge graphs (KGs). The knowledge graph (KG) provides an efficient and clear data structure to store human knowledge. It can be used for knowledge reasoning and retrieval. There have been efforts to enhance the knowledge base of LLMs with KGs [140, 176]. However, it still cannot offer a transparent reasoning process.

With the rapid increase in the size of ML models and the required training resources, green AI [107, 163] has gradually gained attention. Green AI technologies may have the potential to be used in developing chatbots with clearer reasoning processes, smaller model sizes, fewer training resources, and more environmental friendliness. The green learning (GL) methodology [107] has been shown to offer comparable performance with Deep Learning (DL) in many applications. GL methods have much smaller model sizes and lower inference complexities (in terms of FLOPs). They also demand fewer training samples. The GL-based chatbot design may lie in the decomposition of LLMs into two modules: 1) GL-based language models as the interface with respect to users for NLU and NLG tasks; and 2) GL-based KGs as the core for knowledge storage, expansion, search, and reasoning. This high-level idea may guide us to develop a more transparent and scalable chatbot. Biases in GL-based chatbots can be traced, making the implementation of a fair system easier.

6 Conclusion

As an AI system that communicates directly with humans, chatbots have a long development history. They have received special attention in recent years due to the amazing performance of LLM-based chatbots, such as ChatGPT.

Although traditional chatbots are relatively rigid with limited functionalities, their models are smaller and easier to deploy. They have fewer bias and discrimination concerns. They are suitable for small-scale domain-specific applications. Chatbots have become much more powerful with the rapid advancement of LLMs and computing resources in recent years. On the other hand, they have brought controversy about bias and fairness. From development teams to users, every human interacting with an ML-based chatbot has the potential to spread the bias. To design and deploy a fair chatbot system, the development team needs to know social needs, design goals, and actual effects. Overall, the bias and fairness issues of chatbots remain an open problem demanding further research.

References

[1] M. Abbasi, S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "Fairness in representation: quantifying stereotyping as a representational harm", in *Proceedings of the 2019 SIAM International Conference on Data Mining*, SIAM, 2019, 801–9.

[2] R. Abboud, I. Ceylan, T. Lukasiewicz, and T. Salvatori, "Boxe: A box embedding model for knowledge base completion", *Advances in Neural Information Processing Systems*, 33, 2020, 9649–61.

[3] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-muslim bias in large language models", in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, 298–306.

[4] M. Adam, M. Wessel, and A. Benlian, "AI-based chatbots in customer service and their effects on user compliance", *Electronic Markets*, 31(2), 2021, 427–45.

[5] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology", in *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*, Springer, 2020, 373–83.

[6] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications", *Machine Learning with Applications*, 2, 2020, 100006.

[7] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, *et al.*, "Towards a human-like open-domain chatbot", *arXiv preprint arXiv:2001.09977*, 2020.

[8] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial", *IEEE Circuits and Systems Magazine*, 11(2), 2011, 82–108.

[9] P. A. Angga, W. E. Fachri, A. Elevanita, R. D. Agushinta, *et al.*, "Design of chatbot with 3D avatar, voice interface, and facial expression", in *2015 international conference on science in information technology (ICSITech)*, IEEE, 2015, 326–30.

[10] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias", in *Ethics of data and analytics*, Auerbach Publications, 2022, 254–64.

[11] J. Aron, "How innovative is Apple's new voice assistant, Siri?", 2011.

[12] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT generated literature review: Digital twin in healthcare", *Available at SSRN 4308687*, 2022.

[13] R. Baeza-Yates, "Bias on the web", *Communications of the ACM*, 61(6), 2018, 54–61.

[14] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning", *Available at SSRN 4337484*, 2023.

[15] R. S. Baker and A. Hawn, "Algorithmic bias in education", *International Journal of Artificial Intelligence in Education*, 2021, 1–41.

[16] A. Ball-Burack, M. S. A. Lee, J. Cobbe, and J. Singh, "Differential tweetment: Mitigating racial dialect bias in harmful tweet detection", in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 116–28.

[17] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 2018, 423–43.

[18] H. Beattie, L. Watkins, W. H. Robinson, A. Rubin, and S. Watkins, "Measuring and mitigating bias in AI-Chatbots", in *2022 IEEE International Conference on Assured Autonomy (ICAA)*, IEEE, 2022, 117–23.

[19] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias", *arXiv preprint arXiv:1810.01943*, 2018.

[20] J. Bernotat, F. Eyssel, and J. Sachse, "Shape it–the influence of robot body shape on gender perception in robots", in *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, Springer, 2017, 75–84.

[21] J. Bernotat, F. Eyssel, and J. Sachse, "The (fe) male robot: how robot body shape impacts first impressions and trust towards robots", *International Journal of Social Robotics*, 13, 2021, 477–89.

[22] S. S. Biswas, "Potential use of chat gpt in global warming", *Annals of biomedical engineering*, 51(6), 2023, 1126–7.

[23] S. S. Biswas, "Role of chat gpt in public health", *Annals of Biomedical Engineering*, 51(5), 2023, 868–9.

[24] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of" bias" in nlp", *arXiv preprint arXiv:2005.14050*, 2020.

[25] T. Bolton, T. Dargahi, S. Belguith, M. S. Al-Rakhami, and A. H. Sodhro, "On the security and privacy challenges of virtual assistants", *Sensors*, 21(7), 2021, 2312.

[26] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings", *Advances in neural information processing systems*, 29, 2016.

[27] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews", in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, 268–77.

[28] L. Bradeško and D. Mladenić, "A survey of chatbot systems through a loebner prize competition", in *Proceedings of Slovenian language technologies society eighth conference of language technologies*, Vol. 2, sn, 2012, 34–7.

[29] S. Brahnam and A. De Angeli, "Gender affordances of conversational agents", *Interacting with Computers*, 24(3), 2012, 139–53.

[30] P. B. Brandtzaeg and A. Følstad, "Why people use chatbots", in *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, Springer, 2017, 377–92.

[31] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners", *Advances in neural information processing systems*, 33, 2020, 1877–901.

[32] M. Burtell and T. Woodside, "Artificial influence: An analysis of AI-driven persuasion", *arXiv preprint arXiv:2303.08721*, 2023.

[33] J. Cahn, "CHATBOT: Architecture, design, & development", *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*, 2017.

[34] D. M. Cain and A. S. Detsky, "Everyone's a little bit biased (even physicians)", *Jama*, 299(24), 2008, 2893–5.

[35] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots", *Information*, 13(1), 2022, 41.

[36] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science*, 356(6334), 2017, 183–6.

[37] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt", *arXiv preprint arXiv:2303.04226*, 2023.

[38] S. Caton and C. Haas, "Fairness in machine learning: A survey", *arXiv preprint arXiv:2010.04053*, 2020.

[39] A. Chan, "GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry", *AI and Ethics*, 3(1), 2023, 53–64.

[40] K.-W. Chang, V. Prabhakaran, and V. Ordonez, "Bias and fairness in natural language processing", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, 2019.

[41] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design", *International Journal of Human–Computer Interaction*, 37(8), 2021, 729–58.

[42] C. Chen, C. Crivelli, O. G. Garrod, P. G. Schyns, J.-M. Fernández-Dols, and R. E. Jack, "Distinct facial expressions represent pain and pleasure across cultures", *Proceedings of the National Academy of Sciences*, 115(43), 2018, E10013–E10021.

[43] Y. Chen, J. E. Argentinis, and G. Weber, "IBM Watson: how cognitive computing can be applied to big data challenges in life sciences research", *Clinical therapeutics*, 38(4), 2016, 688–701.

[44] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways", *arXiv preprint arXiv:2204.02311*, 2022.

[45] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, *et al.*, "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare", *NPJ digital medicine*, 3(1), 2020, 81.

[46] K. M. Colby, "Modeling a paranoid mind", *Behavioral and Brain Sciences*, 4(4), 1981, 515–34.

[47] K. M. Colby, S. Weber, and F. D. Hilf, "Artificial paranoia", *Artificial intelligence*, 2(1), 1971, 1–25.

[48] P. Costa, "Conversing with personal digital assistants: On gender and artificial intelligence", *Journal of Science and Technology of the Arts*, 10(3), 2018, 59–72.

[49] P. Costa and L. Ribas, "AI becomes her: Discussing gender and artificial intelligence", *Technoetic Arts: A Journal of Speculative Research*, 17(1-2), 2019, 171–93.

[50] K. Crawford, "The trouble with bias", in *Conference on Neural Information Processing Systems, invited speaker*, 2017.

[51] A. C. Curry and V. Rieser, "# MeToo Alexa: how conversational systems respond to sexual harassment", in *Proceedings of the second acl workshop on ethics in natural language processing*, 2018, 7–14.

[52] T. Day, "A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT", *The Professional Geographer*, 2023, 1–4.

[53] M. Delgado-Rodriguez and J. Llorca, "Bias", *Journal of Epidemiology & Community Health*, 58(8), 2004, 635–41.

[54] O. Deryugina, "Chatterbots", *Scientific and Technical Information Processing*, 37, 2010, 143–7.

[55] S. Dev, T. Li, J. M. Phillips, and V. Srikumar, "OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings", *arXiv preprint arXiv:2007.00049*, 2020.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[57] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K.-W. Chang, and R. Gupta, "Bold: Dataset and metrics for measuring biases in open-ended language generation", in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, 862–72.

[58] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification", in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, 67–73.

[59] A. Domnich and G. Anbarjafari, "Responsible AI: Gender bias assessment in emotion recognition", *arXiv preprint arXiv:2103.11436*, 2021.

[60] F. Elsafoury, S. Katsigiannis, and N. Ramzan, "On Bias and Fairness in NLP: How to have a fairer text classification?", *arXiv preprint arXiv:2305.12829*, 2023.

[61] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*, St. Martin's Press, 2018.

[62] F. Eyssel and F. Hegel, "(s) he's got the look: Gender stereotyping of robots 1", *Journal of Applied Social Psychology*, 42(9), 2012, 2213–30.

[63] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "Gender bias in chatbot design", in *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3*, Springer, 2020, 79–93.

[64] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition", *arXiv preprint arXiv:2103.15122*, 2021.

[65] E. Ferrara, "Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies", *arXiv preprint arXiv:2304.07683*, 2023.

[66] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models", *arXiv preprint arXiv:2304.03738*, 2023.

[67] M. Firat, "How chat GPT can transform autodidactic experiences and open education", *Department of Distance Education, Open Education Faculty, Anadolu Unive*, 2023.

[68] R. Fischer and C. Derham, "Is in-group bias culture-dependent? A meta-analysis across 18 societies", *SpringerPlus*, 5(1), 2016, 70.

[69] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of HCI", *interactions*, 24(4), 2017, 38–42.

[70] A. Følstad, C. B. Nordheim, and C. A. Bjørkli, "What makes users trust a chatbot for customer service? An exploratory interview study", in *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*, Springer, 2018, 194–208.

[71] M. Fraiwan and N. Khasawneh, "A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions", *arXiv preprint arXiv:2305.00237*, 2023.

[72] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review", *Neural Networks*, 144, 2021, 187–209.

[73] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination", in *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, 498–510.

[74] I. Garrido-Muñoz, A. Montejo-Ráez, F. Martínez-Santiago, and L. A. Ureña-López, "A survey on bias in deep NLP", *Applied Sciences*, 11(7), 2021, 3184.

[75] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation", *Journal of Artificial Intelligence Research*, 61, 2018, 65–170.

[76] X. Ge, Y. C. Wang, B. Wang, and C.-C. J. Kuo, "Compounding Geometric Operations for Knowledge Graph Completion", in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, July 2023, 6947–65, DOI: 10.18653/v1/2023.acl-long.384, https://aclanthology.org/2023.acl-long.384.

[77] J. Gordon and B. Van Durme, "Reporting bias and knowledge acquisition", in *Proceedings of the 2013 workshop on Automated knowledge base construction*, 2013, 25–30.

[78] J. Gravel, M. D'Amours-Gravel, and E. Osmanlliu, "Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions", *Mayo Clinic Proceedings: Digital Health*, 1(3), 2023, 226–34.

[79] W. Guo and A. Caliskan, "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases", in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, 122–33.

[80] A. Haleem, M. Javaid, and R. P. Singh, "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), 2022, 100089.

[81] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "" I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data", *arXiv preprint arXiv:2212.05856*, 2022.

[82] J.-G. Harms, P. Kucherbaev, A. Bozzon, and G.-J. Houben, "Approaches for dialog management in conversational agents", *IEEE Internet Computing*, 23(2), 2018, 13–22.

[83] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models", in *Proceedings of the European conference on computer vision (ECCV)*, 2018, 771–87.

[84] R. High, "The era of cognitive systems: An inside look at IBM Watson and how it works", *IBM Corporation, Redbooks*, 1, 2012, 16.

[85] M. Hosseini, C. A. Gao, D. M. Liebovitz, A. M. Carvalho, F. S. Ahmad, Y. Luo, N. MacDonald, K. L. Holmes, and A. Kho, "An exploratory survey about using ChatGPT in education, healthcare, and research", *medRxiv*, 2023, 2023–3.

[86] D. Hovy and S. Prabhumoye, "Five sources of bias in natural language processing", *Language and Linguistics Compass*, 15(8), 2021, e12432.

[87] M. B. Hoy, "Alexa, Siri, Cortana, and more: an introduction to voice assistants", *Medical reference services quarterly*, 37(1), 2018, 81–8.

[88] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A survey on conversational agents/chatbots classification and design techniques", in *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33*, Springer, 2019, 946–56.

[89] V. Jain, H. Rai, P. Subash, and E. Mogaji, "The Prospects and Challenges of ChatGPT on Marketing Research and Practices", *Emmanuel, The Prospects and Challenges of ChatGPT on Marketing Research and Practices (March 23, 2023)*, 2023.

[90] A. Janssen, J. Passlick, D. Rodríguez Cardona, and M. H. Breitner, "Virtual assistance in any context: A taxonomy of design elements for domain-specific chatbots", *Business & Information Systems Engineering*, 62, 2020, 211–25.

[91] A. Jiao, "An intelligent chatbot system based on entity extraction using RASA NLU and neural network", in *Journal of physics: conference series*, Vol. 1487, No. 1, IOP Publishing, 2020, 012014.

[92] D. Kaczorowska-Spychalska, "How chatbots influence marketing", *Management*, 23(1), 2019, 251–70.

[93] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammus: A survey of transformer-based pretrained models in natural language processing", *arXiv preprint arXiv:2108.05542*, 2021.

[94] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education", *Learning and Individual Differences*, 103, 2023, 102274.

[95] V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)", in *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, IEEE, 2018, 99–103.

[96] A. Kerlyl, P. Hall, and S. Bull, "Bringing chatbots into education: Towards natural language negotiation of open learner models", in *International conference on innovative techniques and applications of artificial intelligence*, Springer, 2006, 179–92.

[97] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges", *Multimedia tools and applications*, 82(3), 2023, 3713–44.

[98] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks", in *International Conference on Learning Representations (ICLR)*, 2017.

[99] R. F. Kizilcec and H. Lee, "Algorithmic fairness in education", in *The ethics of artificial intelligence in education*, Routledge, 2022, 174–202.

[100] J. Klayman, "Varieties of confirmation bias", *Psychology of learning and motivation*, 32, 1995, 385–418.

[101] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", *arXiv preprint arXiv:1609.05807*, 2016.

[102] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development", *Business Research*, 13(3), 2020, 795–848.

[103] C. Kooli, "Chatbots in education and research: A critical examination of ethical implications and solutions", *Sustainability*, 15(7), 2023, 5614.
[104] N. Kordzadeh and M. Ghasemaghaei, "Algorithmic bias: review, synthesis, and future research directions", *European Journal of Information Systems*, 31(3), 2022, 388–409.
[105] J. E. Korteling, A.-M. Brouwer, and A. Toet, "A neural network framework for cognitive bias", *Frontiers in psychology*, 2018, 1561.
[106] K. M. Kostick-Quenet, I. G. Cohen, S. Gerke, B. Lo, J. Antaki, F. Movahedi, H. Njah, L. Schoen, J. E. Estep, and J. Blumenthal-Barby, "Mitigating racial bias in machine learning", *Journal of Law, Medicine & Ethics*, 50(1), 2022, 92–100.
[107] C.-C. J. Kuo and A. M. Madni, "Green learning: Introduction, examples and outlook", *Journal of Visual Communication and Image Representation*, 2022, 103685.
[108] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness", *Advances in neural information processing systems*, 30, 2017.
[109] K. LaGrandeur, "How safe is our reliance on AI, and should we regulate it?", *AI and Ethics*, 1, 2021, 93–9.
[110] T. Lalwani, S. Bhalotia, A. Pal, V. Rathod, and S. Bisen, "Implementation of a Chatbot System using AI and NLP", *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3*, 2018.
[111] A. Lauscher, G. Glavaš, S. P. Ponzetto, and I. Vulić, "A general framework for implicit and explicit debiasing of distributional word vector spaces", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, 2020, 8131–8.
[112] N. Lee, A. Madotto, and P. Fung, "Exploring Social Bias in Chatbots using Stereotype Knowledge.", in *Wnlp@ Acl*, 2019, 177–80.
[113] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning", *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4), 2019, 297–312.
[114] Y. Li and Y. Zhang, "Fairness of ChatGPT", *arXiv preprint arXiv:2305.18569*, 2023.
[115] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods", *Entropy*, 23(1), 2020, 18.
[116] L. Lucy and D. Bamman, "Gender and representation bias in GPT-3 generated stories", in *Proceedings of the Third Workshop on Narrative Understanding*, 2021, 48–55.
[117] X. Luo, S. Tong, Z. Fang, and Z. Qu, "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases", *Marketing Science*, 38(6), 2019, 937–47.

[118] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey", *Multimedia Tools and Applications*, 80, 2021, 9411–57.

[119] M. McDonnell and D. Baxter, "Chatbots and gender stereotyping", *Interacting with Computers*, 31(2), 2019, 116–21.

[120] R. W. McGee, "Is chat gpt biased against conservatives? an empirical study", *An Empirical Study (February 15, 2023)*, 2023.

[121] P. McNamee, "Language identification: a solved problem suitable for undergraduate instruction", *Journal of computing sciences in colleges*, 20(3), 2005, 94–101.

[122] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning", *ACM Computing Surveys (CSUR)*, 54(6), 2021, 1–35.

[123] P. Meissner and T. Wulf, "The effect of cognitive diversity on the illusion of control bias in strategic decisions: An experimental investigation", *European Management Journal*, 35(4), 2017, 430–9.

[124] V. Mhasawade, Y. Zhao, and R. Chunara, "Machine learning and algorithmic fairness in public and population health", *Nature Machine Intelligence*, 3(8), 2021, 659–66.

[125] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[126] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Algorithmic fairness: Choices, assumptions, and definitions", *Annual Review of Statistics and Its Application*, 8, 2021, 141–63.

[127] R. Mohanani, I. Salman, B. Turhan, P. Rodríguez, and P. Ralph, "Cognitive biases in software engineering: a systematic mapping study", *IEEE Transactions on Software Engineering*, 46(12), 2018, 1318–39.

[128] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach", *AI and Ethics*, 2023, 1–31.

[129] A. Nadeem, B. Abedin, and O. Marjanovic, "Gender Bias in AI: A review of contributing factors and mitigating strategies", 2020.

[130] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models", *arXiv preprint arXiv:2004.09456*, 2020.

[131] C. Nass, Y. Moon, and N. Green, "Are machines gender neutral? Gender-stereotypic responses to computers with voices", *Journal of applied social psychology*, 27(10), 1997, 864–76.

[132] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review", *IEEE access*, 7, 2019, 19143–65.

[133] G. Neff, "Talking to bots: Symbiotic agency and the case of Tay", *International Journal of Communication*, 2016.

[134] G. S. Nelson, "Bias in artificial intelligence", *North Carolina medical journal*, 80(4), 2019, 220–2.
[135] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis", *Applied Sciences*, 9(19), 2019, 4050.
[136] K. K. Nirala, N. K. Singh, and V. S. Purani, "A survey on providing customer and public administration based services using AI: chatbot", *Multimedia Tools and Applications*, 81(16), 2022, 22215–46.
[137] P. Noor, "Can we trust AI not to further embed racial bias and prejudice?", *BMJ*, 368, 2020.
[138] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, 366(6464), 2019, 447–53.
[139] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing", *IEEE transactions on neural networks and learning systems*, 32(2), 2020, 604–24.
[140] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap", *arXiv preprint arXiv:2306.08302*, 2023.
[141] R. B. Parikh, S. Teeple, and A. S. Navathe, "Addressing bias in artificial intelligence in health care", *Jama*, 322(24), 2019, 2377–8.
[142] J. H. Park, J. Shin, and P. Fung, "Reducing gender bias in abusive language detection", *arXiv preprint arXiv:1808.07231*, 2018.
[143] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. R. Bowman, "BBQ: A hand-built bias benchmark for question answering", *arXiv preprint arXiv:2110.08193*, 2021.
[144] J. Paul, A. Ueno, and C. Dennis, "ChatGPT and consumers: Benefits, pitfalls and future research agenda", 2023.
[145] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, 1532–43.
[146] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey", *Science China Technological Sciences*, 63(10), 2020, 1872–97.
[147] M. Qraitem, K. Saenko, and B. A. Plummer, "Bias Mimicking: A Simple Sampling Approach for Bias Mitigation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 20311–20.
[148] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity", *Annals of internal medicine*, 169(12), 2018, 866–72.
[149] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope", *Internet of Things and Cyber-Physical Systems*, 2023.

[150] H. J. Rho, H. Brown, and S. Fremstad, "A basic demographic profile of workers in frontline industries", *Center for economic and policy research*, 7(10), 2020.

[151] P. Rivas and L. Zhao, "Marketing with chatgpt: Navigating the ethical terrain of gpt-based chatbot technology", *AI*, 4(2), 2023, 375–84.

[152] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, E. M. Smith, *et al.*, "Recipes for building an open-domain chatbot", *arXiv preprint arXiv:2004.13637*, 2020.

[153] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in AI", in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, 539–44.

[154] D. Rozado, "The political biases of chatgpt", *Social Sciences*, 12(3), 2023, 148.

[155] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution", *arXiv preprint arXiv:1804.09301*, 2018.

[156] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly, "The Self-Perception and Political Biases of ChatGPT", *arXiv preprint arXiv:2304.07333*, 2023.

[157] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit", *arXiv preprint arXiv:1811.05577*, 2018.

[158] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns", in *Healthcare*, Vol. 11, No. 6, MDPI, 2023, 887.

[159] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions", *SN computer science*, 2(3), 2021, 160.

[160] M. S. Satu, M. H. Parvez, *et al.*, "Review of integrated applications with aiml based chatbot", in *2015 International Conference on Computer and Information Engineering (ICCIE)*, IEEE, 2015, 87–90.

[161] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: Identity, chatbots, and AI", in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, 1–14.

[162] P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do", *Nature Machine Intelligence*, 4(3), 2022, 258–68.

[163] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai", *Communications of the ACM*, 63(12), 2020, 54–63.

[164] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems", in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, 59–68.

[165] S. Seo, “When Female (Male) Robot Is Talking To Me: Effect of service robots’ gender and anthropomorphism on customer satisfaction”, *International Journal of Hospitality Management*, 102, 2022, 103166.

[166] B. A. Shawar and E. Atwell, *A comparison between Alice and Elizabeth chatbot systems*, University of Leeds, School of Computing research report 2002.19, 2002.

[167] B. A. Shawar and E. Atwell, “Fostering language learner autonomy through adaptive conversation tutors”, in *Proceedings of the The fourth Corpus Linguistics conference*, 2007.

[168] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, “Societal biases in language generation: Progress and challenges”, *arXiv preprint arXiv:2105.04054*, 2021.

[169] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, “Towards controllable biases in language generation”, *arXiv preprint arXiv:2005.00268*, 2020.

[170] H.-Y. Shum, X.-d. He, and D. Li, “From Eliza to XiaoIce: challenges and opportunities with social chatbots”, *Frontiers of Information Technology & Electronic Engineering*, 19, 2018, 10–26.

[171] W. M. Si, M. Backes, J. Blackburn, E. De Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang, “Why so toxic? measuring and triggering toxic behavior in open-domain chatbots”, in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, 2659–73.

[172] S. Singh and H. K. Thakur, “Survey of various AI chatbots based on technology used”, in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, IEEE, 2020, 1074–9.

[173] I. Solaiman, “The gradient of generative AI release: Methods and considerations”, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, 111–22.

[174] R. Srinivasan and A. Chander, “Biases in AI systems”, *Communications of the ACM*, 64(8), 2021, 44–9.

[175] S. M. Suhaili, N. Salim, and M. N. Jambli, “Service chatbots: A systematic review”, *Expert Systems with Applications*, 184, 2021, 115461.

[176] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H.-Y. Shum, and J. Guo, “Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph”, *arXiv preprint arXiv:2307.07697*, 2023.

[177] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, “Mitigating gender bias in natural language processing: Literature review”, *arXiv preprint arXiv:1906.08976*, 2019.

[178] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle", in *Equity and access in algorithms, mechanisms, and optimization*, 2021, 1–9.

[179] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning", *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

[180] V. Taecharungroj, ""What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter", *Big Data and Cognitive Computing*, 7(1), 2023, 35.

[181] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis", *arXiv preprint arXiv:2106.15561*, 2021.

[182] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, *et al.*, "Lamda: Language models for dialog applications", *arXiv preprint arXiv:2201.08239*, 2022.

[183] S. A. Thorat and V. Jadhav, "A review on implementation issues of rule-based chatbot systems", in *Proceedings of the international conference on innovative computing & communications (ICICC)*, 2020.

[184] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models", *arXiv preprint arXiv:2302.13971*, 2023.

[185] A. M. Turing, *Computing machinery and intelligence*, Springer, 2009.

[186] D. C. Ukpabi, B. Aslam, and H. Karjaluoto, "Chatbot adoption in tourism services: A conceptual exploration", in *Robots, artificial intelligence, and service automation in travel, tourism and hospitality*, Emerald Publishing Limited, 2019.

[187] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: five priorities for research", *Nature*, 614(7947), 2023, 224–6.

[188] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based Multi-Relational Graph Convolutional Networks", in *International Conference on Learning Representations*, 2020, https://openreview.net/forum?id=BylA_C4tPr.

[189] S. Vasudevan and K. Kenthapadi, "Lift: A scalable framework for measuring fairness in ml applications", in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, 2773–80.

[190] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", *Advances in neural information processing systems*, 30, 2017.

[191] S. Verma and J. Rubin, "Fairness definitions explained", in *Proceedings of the international workshop on software fairness*, 2018, 1–7.

[192] H. Viswanath and T. Zhang, "FairPy: A Toolkit for Evaluation of Social Biases and their Mitigation in Large Language Models", *arXiv preprint arXiv:2302.05508*, 2023.

[193] R. S. Wallace, *The anatomy of ALICE*, Springer, 2009.

[194] C. L. Wang, "Interactive marketing is the new normal", in *The Palgrave handbook of interactive marketing*, Springer, 2023, 1–12.

[195] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models", *arXiv preprint arXiv:2302.07257*, 2023.

[196] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey", *arXiv preprint arXiv:2302.10035*, 2023.

[197] Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo, "An Overview on Generative AI at Scale With Edge-Cloud Computing", *IEEE Open Journal of the Communications Society*, 2023, 1–1, DOI: 10.1109/OJCOMS.2023.3320646.

[198] C. Wei, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "An overview on language models: Recent developments and outlook", *arXiv preprint arXiv:2303.05759*, 2023.

[199] J. Weizenbaum, "Computer power and human reason: From judgment to calculation.", 1976.

[200] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, 9(1), 1966, 36–45.

[201] M. West, R. Kraut, and H. Ei Chew, "I'd blush if I could: closing gender divides in digital skills through education", 2019.

[202] C. Winship and R. D. Mare, "Models for sample selection bias", *Annual review of sociology*, 18(1), 1992, 327–50.

[203] M. J. Wolf, K. Miller, and F. S. Grodzinsky, "Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications", *Acm Sigcas Computers and Society*, 47(3), 2017, 54–64.

[204] H. S. Woods, "Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism", *Critical Studies in Media Communication*, 35(4), 2018, 334–49.

[205] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of ChatGPT: The history, status quo and potential future development", *IEEE/CAA Journal of Automatica Sinica*, 10(5), 2023, 1122–36.

[206] W. Xiong, T. Hoang, and W. Y. Wang, "DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning", in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark: Association for Computational Lin-

guistics, September 2017, 564–73, DOI: 10.18653/v1/D17-1060, https://aclanthology.org/D17-1060.

[207] S. Zabel and S. Otto, "Bias in, bias out–the similarity-attraction effect between chatbot designers and users", in *Human-Computer Interaction. Design and User Experience Case Studies: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part III 23*, Springer, 2021, 184–97.

[208] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification", in *Artificial intelligence and statistics*, PMLR, 2017, 962–70.

[209] M. Żelaszczyk and J. Mańdziuk, "Cross-modal text and visual generation: A systematic review. Part 1—Image to text", *Information Fusion*, 2023.

[210] M. T. ZEMČÍK, "A brief history of chatbots", *DEStech Transactions on Computer Science and Engineering*, 10, 2019.

[211] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion model in generative ai: A survey", *arXiv preprint arXiv:2303.07909*, 2023.

[212] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation", *arXiv preprint arXiv:2305.07609*, 2023.

[213] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods", *arXiv preprint arXiv:1804.06876*, 2018.

[214] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models", *arXiv preprint arXiv:2303.18223*, 2023.

[215] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt", *arXiv preprint arXiv:2302.09419*, 2023.

[216] J. Zhou, H. Müller, A. Holzinger, and F. Chen, "Ethical ChatGPT: Concerns, Challenges, and Commandments", *arXiv preprint arXiv:2305.10646*, 2023.

[217] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot", *Computational Linguistics*, 46(1), 2020, 53–93.

[218] J.-J. Zhu, J. Jiang, M. Yang, and Z. J. Ren, "ChatGPT and environmental research", *Environmental Science & Technology*, 2023.

[219] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Exploring ai ethics of chatgpt: A diagnostic analysis", *arXiv preprint arXiv:2301.12867*, 2023.

[220] D. Zumstein and S. Hundertmark, “CHATBOTS–AN INTERACTIVE TECHNOLOGY FOR PERSONALIZED COMMUNICATION, TRANSACTIONS AND SERVICES.”, *IADIS International Journal on WWW/ Internet*, 15(1), 2017.