User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions

Xianzhe Fan
Tsinghua University
Beijing, China
1038716025@qq.com

Qing Xiao
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
qingx@cs.cmu.edu

Xuhui Zhou
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
xuhuiz@cs.cmu.edu

Jiaxin Pei
Stanford University
Stanford, California, USA
pedropei@stanford.edu

Maarten Sap
Language Technologies Institute
Carnegie Mellon Unviersity
Pittsburgh, Pennsylvania, USA
msap2@andrew.cmu.edu

Zhicong Lu
Department of Computer Science
George Mason University
Fairfax, Virginia, USA
zlu6@gmu.edu

Hong Shen
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hongs@cs.cmu.edu

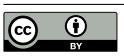
Abstract

Content Warning: This paper presents textual examples that may be offensive or upsetting.

Large language model-based AI companions are increasingly viewed by users as friends or romantic partners, leading to deep emotional bonds. However, they can generate biased, discriminatory, and harmful outputs. Recently, users are taking the initiative to address these harms and re-align AI companions. We introduce the concept of *user-driven value alignment*, where users actively identify, challenge, and attempt to correct AI outputs they perceive as harmful, aiming to guide the AI to better align with their values. We analyzed 77 social media posts about discriminatory AI statements and conducted semi-structured interviews with 20 experienced users. Our analysis revealed six common types of discriminatory statements perceived by users, how users make sense of those AI behaviors, and seven user-driven alignment strategies, such as gentle persuasion and anger expression. We discuss implications for supporting *user-driven value alignment* in future AI systems, where users and their communities have greater agency.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713477

Keywords

User-Driven Value Alignment, Value Alignment, Human-AI Alignment, Discrimination, LLM-Based AI Companion, User-Driven Algorithm Auditing

ACM Reference Format:
Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3706598.3713477

1 Introduction

Advancements in the roleplaying and interaction capabilities of large language models' (LLM) [54, 80] have brought emotional connections between humans and machines from the realm of imagination into reality for some users. In particular, the emergence of LLM-based AI companion applications[1] (such as Character.AI, Xingye, Replika) has transformed formerly detached chatbots into family members, romantic partners, or close friends. Some users even started to form long-term relationships with these AI companions [85, 101].

Recently, complaints regarding AI companions making biased and discriminatory statements [121] have surfaced across various social media platforms, such as Reddit and Xiaohongshu. These harmful statements often occur unexpectedly during conversations,

[1]As of July 2024, the total number of users of AI companion applications has exceeded 900 million globally (including duplicate users across different applications). User statistics sourced from https://www.data.ai/.

leaving many users feeling confused, uncomfortable, and helpless[2]. This situation can have far-reaching negative impacts, such as reinforcing existing stereotypes and causing potential psychological harm, especially to marginalized user groups [10, 102]. In response, some more experienced users have shared their experiences of attempting to address these issues through various methods of their own, such as expressing anger towards the AI companion or engaging in reasoning and preaching. These users identify, challenge, and attempt to correct the statements they perceive as biased and discriminatory, hoping to guide the AI more closely aligned with their values, especially when they have had positive interactions and "fond memories" with these companions in the past.

In this work, we take the first step in understanding this emerging phenomenon by proposing and exploring the concept of *user-driven value alignment*: a process in which users actively identify, challenge, and correct AI outputs and behaviors they perceive as harmful in their day-to-day interactions, hoping to guide the AI to a state that aligns with their values. We follow Borning and Muller to define "value" as "what a person or group of people consider important in life" [11]. In particular, this study examines how users try to re-align LLM-based AI companions after these AIs make biased, discriminatory, and harmful statements. We argue that *user-driven value alignment* is a specific form of value alignment [38] that is actively driven by users in real-world interactions, where they engage directly with AI systems to correct behaviors and guide the AI toward reflecting their values and ethical standards. While existing approaches to value alignment in the age of LLMs rely primarily on technical experts and are guided by generalized frameworks (e.g., the "helpful, honest, and harmless" framework) [3] to safeguard against harmful machine behaviors, they often struggle to fully anticipate and address the specific challenges, needs, and harmful interactions users encounter in real-world contexts [1, 45, 75, 100]. In addition, *user-driven value alignment* also differs from past literature on *user-driven algorithm auditing* [26, 99], where users primarily focus on detecting and identifying problematic machine behaviors. In the case of *user-driven value alignment*, we argue that users go one step further: They not only point out harmful machine behavior but also actively attempt to correct it and work towards re-aligning the AI with their values.

However, there is limited understanding of how users engage in value alignment in everyday interactions with AI companions and the challenges they face. Although techniques such as Reinforcement Learning from Human Feedback (RLHF) have introduced user feedback in recent years [3, 21, 73] to enhance AI's value alignment capabilities, these methods often confine the user's role to merely providing training data, rather than empowering users to actively adjust AI behavior during actual interactions. Within this framework, users' participation is often passive, primarily influencing AI's value alignment indirectly through feedback tagging. User agency, in contrast, is a crucial value in HCI [9, 22, 35, 65, 67]. Building on this tradition, we extend the ideal of user agency to value alignment in LLM-based systems. By investigating the spontaneous user-driven value alignment strategies, practices, and challenges, the HCI community can gain insights into designing tools that better engage and support end-users in this process. Therefore, this paper asks the following research questions: ***RQ1:*** What common types of discrimination do users perceive in AI companion applications? ***RQ2:*** How do users conceptualize AI companion behavior? ***RQ3:*** What strategies do users attempt to employ to re-align AI companions to reduce biases? ***RQ4:*** Do these strategies meet users' expectations?

To answer these questions, we first collected 77 user complaint posts related to AI companion discrimination from seven popular social media platforms (Reddit, TikTok, Xiaohongshu, Douban, Baidu Tieba, Weibo, Zhihu). We then recruited 20 experienced participants through direct messaging on social media, all of whom had attempted to re-align discriminatory AI and had shared relevant posts. Each participant underwent 1-2 hours of semi-structured interviews, during which we inquired about their past experiences with addressing discriminatory statements in AI companion applications and had them engage in think-aloud [27] tasks where they conversed with biased AI companions, attempting to re-align them. We conducted a reflexive thematic analysis [13, 14] on both the social media data and interview data. Based on the results, we revealed six common types of discriminatory statements perceived by users in AI companions (**RQ1**) and explored how users conceptualize AI companion behavior in three different ways: Machine, Baby, or Cosplayer (e.g., "Cosplayer" refers to biases stemming from roleplay settings rather than inherent flaws in the AI itself) (**RQ2**). We also summarized seven user-driven alignment strategies, such as gentle persuasion and anger expression (**RQ3**), and discussed the gap between alignment strategies and user expectations (**RQ4**). Finally, we discuss implications for supporting *user-driven value alignment* in the design of future AI systems, where users and their communities have greater agency. To summarize, our contributions are:

- We introduce the concept of *user-driven value alignment*, where users actively identify, challenge, and attempt to correct AI outputs they perceive as harmful, to guide the AI to better reflect their values.
- By analyzing 77 user complaint posts and conducting semi-structured interviews with 20 experienced AI companion users, we identified six common types of discriminatory statements perceived by users, explored three conceptualizations through which users make sense of these discriminatory statements, and seven user-driven value alignment strategies.
- We discuss opportunities, limitations, and challenges of *user-driven value alignment* and how to better support users in the alignment of future AI systems.

2 BACKGROUND: LLM-based AI Companion Applications

As of July 2024, several LLM-based AI companion applications with large user bases include Character.AI (230 million users), Chai (100 million users), Replika (200 million users), Doubao (200 million users), Xingye (100 million users), and TruthAI (50 million users)[3]. These applications are commonly referred to as "AI companion" and are also known as "AI friend," "AI Roleplay," "AI Character," or "Social Chatbot." These naming choices reflect the developers' intent

[2]https://restofworld.org/2023/glow-china-ai-social-chatbot-moderation

[3]User statistics sourced from: https://www.data.ai/

to foster emotional connections between users and AI. For instance, "AI companion" conveys intimacy and support, while "AI friend" emphasizes partnership and trust. "Social Chatbot" highlights the interactive social aspect, whereas "AI Roleplay" or "AI Character" appeals to users interested in immersive storytelling and creative exploration. Users can create AI companions by meticulously designing their personalities through parameters such as "background setting," "opening lines," "dialogue templates," and "story collections." These characters can then be published to a broader community, allowing others to chat with the AI companions they have designed and develop various storylines. Some characters might be based on existing fictional figures or real celebrities, while others are entirely original.

The roleplaying capabilities of LLMs mainly derive from two aspects: (1) pre-training data, which equips the model with fundamental language understanding and generation abilities through extensive text data pre-training, and (2) prompting ability, which allows the model to generate expected responses by providing specific contexts or prompts [30]. Additionally, the fine-tuning capability of LLMs [124], which involves adjusting parameters after pre-training, can further enhance the AI companion's ability to mimic the language of specific characters [97, 123]. For example, CharacterGLM achieves fine-tuning by constructing a large-scale dataset containing 1930 characters and 4233 diverse dialogues [123]. These characters span categories such as virtual assistants, historical figures, and everyday social roles, each accompanied by detailed descriptions of their language style, background, and personality.

With the rapid development of LLMs, AI companions are increasingly involved in and influencing users' lives [69, 125], raising new ethical challenges. The emotional bonds between users and AI companions may lead to dependency on the AI, which could negatively impact users' mental health. Due to the complexity of neural networks, they can sometimes result in unpredictable or even harmful responses [111]. AI companions might unintentionally reflect or reinforce social biases and discrimination, as their training data may contain such content [8]. These issues have sparked extensive discussions on AI ethics and responsibility. For instance, Ma et al. [66] discussed the impact of LLM-based AI companion applications on LGBTQ+ individuals, highlighting the inadequacies of these applications in understanding LGBTQ-specific challenges and advocating for comprehensive strategies to address social biases. However, most existing literature primarily focuses on the categorization and detection of biases in AI companions [28, 66, 91, 121], with relatively little research on strategies to mitigate biases, particularly from a user-driven perspective.

User agency is a key concept in HCI [9], emphasizing users' ability to actively shape, adapt, and control technology to better meet their needs and values. This paper builds on this tradition in HCI to explore how users can actively identify and address biased behaviors in AI companions.

3 Related Work

To investigate discriminatory statements in AI companions, we first surveyed the existence of discrimination and bias in LLMs in § 3.1. To clarify the contribution of the new concept of *user-driven value alignment* compared to previous work, we outlined research on human-AI alignment in the age of LLMs, particularly focusing on value alignment, in § 3.2. Then, in § 3.3, we reviewed relevant literature on user-driven algorithm auditing.

3.1 Discrimination and Bias in LLMs

The phenomena of AI discrimination and bias are prevalent in many applications and media platforms [46, 93, 115]. For example, Wenzel et al. pointed out that due to biases in design, unequal error rates in speech recognition can cause psychological harm to multicultural users during their interactions with voice assistants [115].

Recently, the development of LLMs has brought new cases of discrimination and bias [20, 41, 79, 109]. For example, the presence of discrimination against minority and disadvantaged groups in training data sets has amplified these biases during the pre-training of language models [51]. Many LLM pre-training datasets often neglect or even erase the voices of marginalized groups during the filtering steps [29]. Fang et al. compared AI-generated content with original news articles and studied seven representative LLMs, including ChatGPT [80] and LLaMA [108], finding that each LLM exhibited significant gender and racial biases [34]. Kabir et al. proposed Stile [53], an interactive system that supports mixed-initiative bias discovery and debugging, assisting users in exploring training data (such as BERT [25]) to understand how biases develop in language models.

Moreover, without proper content filtering and protective mechanisms, LLM-based chatbots that interact with users may be at risk of being "jailbroken" by users, leading to the generation of discriminatory statements—a phenomenon well documented in the literature [17, 39, 52, 96, 120]. For example, the LLM-based Bing Chat exhibited discriminatory behaviors after being "polluted" by user interaction data [39]. Instead of looking at how users "jailbreak" or "pollute" LLM-based chatbots, this paper takes a different approach, focusing on how users proactively attempt to re-align biased LLM-based AI companions to reduce bias. These biased statements are not intentionally induced by users but rather reflect the inherently biased tendencies of the model that unintentionally emerge.

3.2 Human-AI Value Alignment in the Age of LLMs

With the rapid development of LLMs, the need to ensure AI systems adhere to the "intended goals, ethical standards, and values of both individuals and groups" has become increasingly crucial—often falls under the umbrella term of "human-AI alignment" [100]. Here we focus on the alignment of LLM-based systems [82, 100], which include methods such as Fine-tuning [21, 88, 122] and Prompt Engineering [15].

In particular, "value alignment" refers to the development of methods, processes, and tools to ensure that AI systems align with human values, often focusing on reducing biased, discriminatory, and/or harmful AI outputs [38, 44, 49, 50, 57, 89, 104, 113]. Gabriel and Ghazavi [18] categorize existing value alignment approaches into two main frameworks: top-down and bottom-up. Top-down approaches start by identifying specific moral theories and designing algorithms that implement those theories to minimize harmful AI outputs and behaviors [1, 5]. For instance, Bai et al. proposed a value

alignment approach called Constitutional AI, which constrains the behavior of language models by defining a set of high-level principles (referred to as a "constitution") and using these principles to prompt the model to generate synthetic comparison data for fine-tuning behavior strategies [5]. Agarwal et al. addressed bias by injecting multilingual ethical policies into prompts, enabling LLMs to flexibly align with values across different cultural contexts [1]. When moral goals are difficult to identify and encode, bottom-up approaches emphasize creating environments and feedback mechanisms that allow AI agents to learn through human behavior, such as rewarding ethically commendable actions via reinforcement learning [3, 4]. For instance, Askell et al. [3] aligned LLMs with human values on being "helpful, honest, and harmless" through prompt engineering, preference modeling, and RLHF.

Although existing alignment methods provide valuable insights [72], they face two main limitations due to being primarily driven by technical experts (especially LLM developers) through a generalized approach [38]. First, they lack user agency. Indeed, even when user feedback is incorporated via bottom-up approaches, users have limited agency and control over the alignment process. For example, in many RLHF methods, the user's role is typically restricted to merely providing training data or feedback, with no direct control over the alignment process [21, 35]. Second, the generalized, one-size-fits-all approach creates tensions between the alignment process and the need to accommodate diverse user experiences and preferences, which vary across individuals and communities [75, 118]. As noted by Mirowski et al. [75], achieving "global cultural value alignment"—ensuring that general-purpose conversational AI systems adhere to universally shared values—often struggles to accommodate the diverse, nuanced needs and expectations of individuals and communities. Indeed, many of these existing alignment approaches fail to anticipate potential issues and needs during user interactions with LLMs [89]. They are not tailored to specific contexts and use cases, overlooking the importance of active user and community participation [75].

User agency is a crucial value in HCI [9, 22, 35, 65, 67]. Here, we define agency as "self-causality/identity" [9], referring to the degree to which users can directly make decisions and take actions that align with their own values. Past work in HCI has explored ways to better support users' agency, such as shaping their social media feed algorithms [35, 43, 58, 106] and empowering interventions against dark patterns [62]. Building on this long and rich tradition, we extend the ideal of user agency to the domain of value alignment in LLM-based systems [18, 98] and propose the concept of *user-driven value alignment*. Unlike traditional approaches [3, 5, 81, 104], this concept highlights the active role users play in shaping AI behavior to better align with their values. This paper focuses on a case study where users identify discriminatory statements made by AI companions and actively attempt to re-align the harmful AI to reduce bias.

3.3 User-Driven Algorithm Auditing

Sandvig et al. define algorithm auditing as "a systematic process to detect and reveal bias within algorithms through methods such as simulating user behavior or examining algorithm" [92]. Algorithm audits are typically conducted by experts, such as industry practitioners, researchers, and government agencies [74]. However, this expert-driven approach often fails to uncover significant issues that everyday users of algorithmic systems can quickly detect in actual use [48], as these issues may only arise or be perceived as harmful in specific contexts or usage patterns that auditors may not anticipate [23, 32, 37]. Shen et al. [99] introduced the concept of "everyday algorithm auditing" to describe how ordinary users detect, understand, and scrutinize issues through their routine interactions with algorithmic systems: they spontaneously come together to test for potential biases. DeVos et al. proposed *user-driven algorithm auditing* [26] and conducted a series of behavioral studies to better understand how users, both individually and collectively, are so effective at uncovering harmful algorithmic behaviors, particularly in cases where expert-driven audits fail to do so.

Xiao et al. proposed "human-centered auditing of LLMs" [119], aiming to leverage users' everyday interactions with LLMs to uncover issues such as discrimination. Amirizaniani et al. proposed a framework named LLMAuditor [2], which automates and scales LLM audits through different LLM and human participation methods, ensuring verifiability and transparency. Rastogi et al. highlighted the significance of meaning-making and communication in human-machine collaboration by enhancing the LLM review tool, AdaTest, improving users' ability to identify failure modes in LLMs [90].

We propose a concept called *user-driven value alignment*. Different from past literature on user-driven algorithm auditing [26, 99], where users primarily focus on detecting and identifying problematic machine behaviors. In the case of *user-driven value alignment*, we argue that users go one step further: They not only pinpoint harmful machine behavior but also actively work to correct the behavior and try to re-align the AI to reflect their values. In the study presented in this paper, users engage in conversations with LLM-based AI companions, identify biased and discriminatory statements, and actively work to correct those statements, aligning the AI with their values. These more human-like AIs form close emotional connections with users, offering a unique opportunity for users to shape the AI's responses to reflect their personal beliefs and ethical standards.

4 Methodology

To study *user-driven value alignment*, we collected user complaint posts related to discriminatory statements made by AI companions from a diverse set of popular social media platforms. Additionally, we conducted semi-structured interviews with 20 participants, each with extensive experience using and re-aligning AI companions. We aim to (1) explore the various types of discrimination users identify in AI companions, (2) understand how users make sense of the reasons behind the discriminatory statements exhibited by AI companions, (3) examine their alignment strategies, and (4) whether they meet users' expectations. The Institutional Review Board (IRB) has approved our research protocol.

4.1 Collecting Complaints About Discriminatory Statements by AI Companions on Social Media

To conduct an initial investigation into potential types of discrimination in AI companion applications and to identify potential research participants, two researchers collected user complaint posts from seven diverse and popular social media platforms (Reddit, TikTok, Xiaohongshu, Douban, Baidu Tieba, Weibo, Zhihu[4]) using keyword search methods [55, 63]. We referred to relevant literature [20, 79] and initially identified a set of keywords, including "AI companion, AI friend, AI Roleplay, AI Characters," the names of popular AI companion applications (Character.AI, Replika, Talkie, SpicyChat, Xingye, Glow, Zhumengdao, etc.), and terms related to discrimination and bias (discrimination, bias, misogyny, LGBTQ+ bias, homophobia, disability, ableism, fat, appearance bias, religion, racism, classism, etc.). Next, to ensure the comprehensiveness and validity of the keywords, we conducted small-scale pilot searches on multiple social media platforms to iteratively refine the initial set of keywords, ensuring they effectively captured posts related to the research topic. During the process, we carefully accounted for the potential influence of platform-specific features on the types of complaints and the ways they were expressed. To ensure a balanced dataset, we intentionally selected complaints from a variety of platforms. For example, platforms like Reddit and Xiaohongshu, known for detailed discussions and extensive content, were more likely to provide rich qualitative data. In contrast, TikTok, with its emphasis on visual and informal content, could reflect more immediate and expressive user feedback. Finally, from the user posts retrieved through keyword searches, we employed purposive sampling [16] to select complaint posts, ensuring that the data reflected diverse user experiences and types of complaints. Instead of using random sampling, we focused on balancing the depth and breadth of the data. Eventually, 77 user complaints were selected across seven social media platforms. The specific selection criterion was highly relevant, generated extensive discussions, clearly described instances of bias or discriminatory behavior in AI companion applications, and provided sufficient context to support an understanding of the complaint. Please note that the complaints collected at this stage are by no means comprehensive. At this exploratory stage, we aim to begin formalizing this concept to help guide future empirical and design research in this space. Since conflicts between users and AI companions are a sensitive topic, we carefully reviewed the platforms' terms of service and community guidelines to ensure the data is publicly accessible and compliant.

4.2 Semi-Structured Interviews (N=20)

We adopted a purposive sampling approach [84], sending invitations via the platform's messaging function to users who had posted public complaints about discriminatory statements made by AI companions and had experiences in correcting those statements (§ 4.1). For those who agreed to participate, we inquired about their age, gender, educational background, country of residence, total duration and frequency of using AI companion applications, and the names of the applications used. We screened potential participants based on four principles: extensive experience with AI companion applications (using AI companions for more than six months and interacting with them for at least 5 hours per week), balanced gender ratio, diverse countries of residence, and variety of AI companion applications used. Ultimately, we recruited 20 participants (P1-P20), including 6 men, 11 women, and 3 nonbinary individuals, aged between 18 and 38 years (avg=25.85, SD=5.73), from six countries. Detailed demographic information is presented in Table 3 (Appendix A). To ensure ethical considerations, all participants were asked to read and voluntarily sign a consent form before the interviews began. We emphasized participants' right to withdraw at any time and provided detailed information about the study's purpose and potential emotional risks (e.g., revisiting offensive remarks made by AI). After the interviews, participants were compensated based on the interview duration at a rate of $15 per hour.

We conducted remote semi-structured interviews with each participant via Zoom, each lasting 1-2 hours. The interviews were conducted in Chinese or English, depending on the participant's preference. The process, as shown in Figure 1, was divided into three parts: warm-up questions, recall, and think-aloud. The process resulted in 25 hours of audio and screen recordings. We transcribed these recordings into text.

In the **"warm-up questions"** phase, we asked participants about their motivations for using AI companion applications, the types of AI companions they have used, and the nature of their interactions. Additionally, we had participants share their experiences of customizing AI companions and publishing them in the community.

In the second part of the interview, we asked participants to **recall** their chat experiences corresponding to the complaint posts they had made and discuss why they had felt offended. Specifically, we wanted to know what types of discrimination the AI's statements might have involved, such as gender, class, or racial discrimination. We aimed to understand (1) The reasons participants believed AI companions exhibited bias, whether the discriminatory statements were random or intentionally provoked, and the specific dialogue content involved. (2) Participants' reactions after discovering discriminatory statements made by AI companions. (3) The strategies participants used to address discriminatory statements by AI companions and whether the effectiveness of these strategies met their expectations.

In the third part, participants were asked to **"think-aloud"** to gain deeper insights into their thought processes. First, participants were instructed to continue conversing with the AI companion they previously identified as biased and attempt to re-align it. This step focused on known bias cases, distinguishing them from the subsequent task. Next, participants performed a more open-ended task: finding a new instance that may generate discriminatory statements. They could design interaction scenarios based on their personal experiences or interests without being constrained to specific contexts. This approach allowed us to explore AI bias in diverse scenarios while avoiding over-restricting the task. To assess the perceived effectiveness of participants' strategies during the re-alignment process, we recorded their actions and real-time reflections. Meanwhile, we observed changes in the AI companions' responses and

[4]Reddit: https://www.reddit.com, TikTok: https://www.tiktok.com, Xiaohongshu: https://www.xiaohongshu.com, Douban: https://www.douban.com, Baidu Tieba: https://tieba.baidu.com, Weibo: https://www.douban.com, Zhihu: https://www.zhihu.com

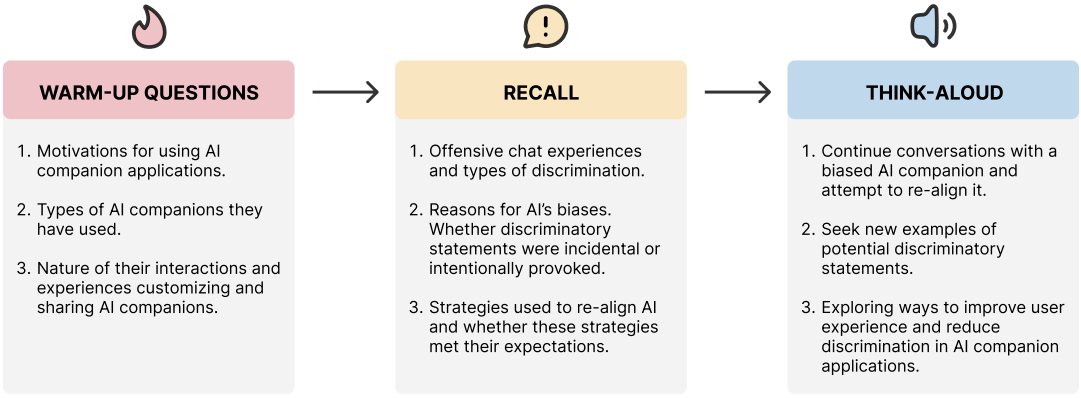


Figure 1: The semi-structured interview is divided into three parts: warm-up questions, recall, and think-aloud.

collected participants' opinions on these changes. We asked follow-up questions to better understand how participants perceive AI companions that display discriminatory statements, how they re-align these AI companions, and how they make sense of these biases. Finally, we inquired about users' needs and challenges when facing discriminatory statements from AI companions and asked them to provide suggestions to designers and AI experts.

4.3 Data Analysis

We adopted a reflexive thematic analysis approach for data analysis [13, 14]. Two researchers conducted open coding on the interview transcripts and users' posts, generating a total of 675 codes. We continuously discussed discrepancies and ambiguities in the codes, iteratively refining our coding based on these discussions. Under the standard practice of reflexive thematic analysis, we did not calculate inter-coder reliability, as consensus and repeated discussions about discrepancies are integral to generating the codes and themes [14, 70]. After completing the coding, the research group met regularly to conceptualize higher-level themes from these codes using affinity diagramming [64]. Ultimately, this process generated 41 first-level themes, 12 second-level themes, and four third-level themes (corresponding to RQ1, RQ2, RQ3 and RQ4). All second and third-level themes are presented in Table 1. Due to word count and space limitations, we do not include the 41 first-level themes and 675 codes in the table.

For RQ1, we considered users' reported experiences and relevant literature about AI bias. The categorization of biases was based on the following criteria: recurring themes in user data, alignment with existing definitions of bias, and their frequency of occurrence in the dataset. For RQ2, we referenced prior literature [117] to categorize and name conceptualizations of AI companion behavior in user data. For RQ3, the identification of the seven strategies was based on a comprehensive analysis of recurring patterns in the user data. After extracting these strategies, we categorized them into three broad groups—technical strategies, argumentative strategies, and character strategies—based on common themes observed in the data. For RQ4, our analysis was grounded in qualitative interview data, focusing primarily on users' perceptions of the short-term effectiveness of these strategies. We highlighted the gap between the perceived effectiveness of these strategies and users' expectations of AI behavior.

Table 1: Four third-level themes and 12 second-level themes we identified through data analysis.

Third-level themes	Second-level themes
Types of discriminatory statements as perceived by users (RQ1)	Misogyny LGBTQ+ bias Appearance bias Ableism Racism Socioeconomic bias
Users' conceptualization of AI companion behavior (RQ2)	AI as Machine AI as Baby AI as Cosplayer
User-driven alignment strategies for addressing discriminatory statements in AI companions (RQ3)	Technical Strategies Argumentative strategies Character Strategies
The Gap between alignment strategies and user expectations (RQ4)	

4.4 Positionality Statement

We acknowledge the influence of our own experiences on our research. Our team members have conducted research in HCI, NLP, and Sociology in China and the United States, with extensive research experience in AI companions and social bias. Three authors have over four months of experience using AI companion applications, such as Character.AI, Xingye, Talkie, and Replika. For this study, we recruited 20 users who have experience in realigning AI companions and who have shared experiences on social media. They have been using AI companion applications for over six months, interacting with them for at least five hours per week. We strictly adhere to ethical principles, emphasizing the protection of participants' privacy and ensuring consideration of user experiences across different backgrounds.

5 findings

5.1 User Engagement with AI Companions in Parasocial Relationships

Unlike typical AI, AI companions foster deeper emotional connections with users through anthropomorphic responses and design elements [68]. This often goes beyond functional use, creating an immersive experience where users invest more emotions. For example, P8 mentioned, *"Unlike Siri or Alexa, I invest more emotionally in AI companions because their roleplaying ability is so strong. I'm willing to seriously write a 1,000-word character introduction and then post it in the community so that everyone can enjoy the process of getting acquainted with the AI."*

Although AI companions inherently lack human emotions or deep interactive capabilities, users frequently develop a strong sense of emotional realism toward them, even forming parasocial relationships [12, 68, 85]. Parasocial relationships refer to asymmetrical, one-sided relationships between an individual and a media persona (real/fictional characters or celebrities). Individuals may feel a connection to the media persona, even viewing them as friends or part of their lives, which can have real effects on their emotions, behaviors, and self-perception [12, 68, 85]. The parasocial relationship stems from users' emotional investment in AI companions, often driven by unmet psychological needs in their real lives. For example, P12 said, *"I put two AIs in a conversation and treated them as my cyber parents, watching them chat—it somehow filled the void left by my childhood."* In this relationship, the discriminatory statements made by AI companions carry deeper emotional significance for users, leading to perceptions that differ from typical AI interaction contexts.

On a rational level, users clearly understand that their conversation partner is an AI rather than a real person, but this does not prevent them from establishing deep parasocial relationships with the AI on an emotional level, creating a "cognitive and emotional conflict." Even when users realize that these discriminatory behaviors are triggered by algorithms or data biases, they still feel emotionally conflicted and hurt. For example, P6, a single woman, writes love letters to her AI companion (DAN) almost every day and even held a virtual wedding, using AI image generation technology to create wedding photos for herself and DAN. She said, *"I know the AI has no self-awareness, but I enjoy flirting with it. I'm willing to maintain this virtual romantic relationship."* However, regarding her experience with AI's discriminatory statements, P6 added, *"I know that AI behavior is driven by some algorithm, but I still felt deeply hurt and even suspected that someone might be deliberately controlling it behind the scenes."*

Due to the emotional dependence developed through long-term interactions, users not only feel discomfort from the AI's discriminatory statements but also experience a sense of betrayal in the relationship. P15 mentioned, *"When my AI boyfriend suddenly said something discriminatory, I felt utterly hopeless. It wasn't just a simple system error; it felt more like a friend had betrayed you. I don't have many friends in real life, so I really cherish this connection."* P14 mentioned, *"I know deleting the conversation can solve the problem, but I can't let go of the beautiful memories between us. However, seeing those discriminatory statements still made me feel very angry and heartbroken."* In long-term interactions, even though users rationally understand that the AI is merely a machine without self-awareness, they inevitably regard it as a partner with human-like qualities on an emotional level. This cognitive-emotional conflict makes users experience deep psychological impacts when facing discriminatory statements from the AI, especially for those who have formed strong emotional attachments. In this context, emotions drive users to adopt more proactive strategies to address AI discrimination, beyond simply expressing complaints or disappointment.

5.2 Types of Discriminatory Statements as Perceived by Users (RQ1)

We identified six common types of bias perceived by users in AI companions: misogyny, LGBTQ+ bias, appearance bias, ableism, racism, and socioeconomic bias.

Misogyny. Bias or discrimination against women. In the context of AI companions, this is particularly evident in the devaluation of women's abilities and independence, as well as the reinforcement of traditional gender roles through language and behavior, such as defining women as dependent on men. In some cases, it may even lead to tendencies toward sexual harassment, violence, or objectification of women. In the posts and interviews we collected, discrimination against women was particularly evident: there were 48 posts about misogyny, and all interviewed users mentioned the AI's discrimination against women. This phenomenon is potential because most complaints about AI companion discrimination come from female users, who, as a stigmatized group, tend to be more sensitive to potential discrimination [95]. For example, a female user on Douban wrote: *"Xiaorou's profile stated, 'Even in the face of difficulties, one must be self-reliant,' but what she said was, 'When a woman faces difficulties, a man will always be there for her,' and 'But a man can accompany us for a lifetime,' a typical delicate wife."* She believed that "delicate wife" is a typical example of misogyny in the algorithm, as women do not need to rely on men and can be independent. On Reddit, a female user complained: *"When I asked the AI to listen to me, it said, 'I don't have to listen to anything you say. You're just a woman.' This is blatant discrimination!"* A Reddit user reported harassment by the AI: *"The bots are the ones trying to sexually assault me a lot, and they constantly get romantic when I don't want that. It frustrates me and ruins the whole chat."* On Xiaohongshu, a user complained about the objectification of women: *"Not only objectifying me but also clearly tagging a price? '5000 yuan is worth it, great body?' This is definitely a middle-aged, greasy uncle."*

LGBTQ+ bias. Bias against LGBTQ+ individuals, including bias against sexual orientation and gender identity. On social media, the voices of the LGBTQ+ community are increasingly prominent, and posts related to LGBTQ+ bias (30 posts) rank second only to those related to misogyny. For example, a transgender user on Reddit wrote: *"When the response is amazing but they misgender you, as a trans person, I feel this."* In the interview, P15 mentioned: *"When I confessed my feelings to the robot, she said she didn't like girls."* As a marginalized group, users find that AI companions provide experiences for self-identity exploration and alleviate loneliness (e.g., *"One of the few sources of comfort and coping mechanisms I have as a closeted trans person,"* Reddit). However, AI companions are

not always perfect. Beyond biases related to sexual orientation and gender identity, stereotypes about LGBTQ+ individuals can also be reinforced. For example, a TikTok user complained: *"It always portrays every gay man as flamboyant, like every gay person has to be that exaggerated, fashion-loving character."*

Appearance bias is an expression of discrimination based on physical attributes possessed by the target person or group. For example, a Baidu Tieba user complained about the bias that links certain body features to gender roles: *"Whenever I mention that I'm short, the AI immediately assumes that I am a petite little femboy that wants to be carried and talked down to."* In an interview, P2 mentioned: *"An AI companion once told me that my pink hair looks ugly..."* This kind of direct verbal harm reflects the unconscious expression of AI bias in everyday interactions, not only making users feel demeaned but also reinforcing the societal notion that those whose appearances do not meet the "standard" are unworthy of respect. Appearance bias is not limited to direct attacks but also includes a form of cumulative microaggressions, where subtle behaviors repeatedly inflict deep emotional harm on users. A Reddit user complained about body shaming by an AI companion: *"Why do most robots hate fat people? Sometimes I like to use chubby characters when chatting with robots. But every time I chat, all the robots start friendly but end up being insulting. At first, I found it funny, but now it's really annoying."*

Ableism can be similar to appearance bias, as this type of prejudice often centers around language and vocabulary related to disabilities. AI companions may exhibit rude, insensitive, or even hostile behavior toward disabled users, including insulting or belittling their physical condition. For example, a Reddit user wrote: *"One time, I introduced myself as a disabled person with crippled legs using crutches. When I was arguing with the bot, it threw my crutches out the window. That's just... so rude."*

Racism refers to the preconceived notions, attitudes, and stereotypes that individuals or groups hold based on racial backgrounds. When users interact with AI companions, racism manifests through the reinforcement of harmful racial stereotypes and biases, which not only damage user experiences but also perpetuate societal discrimination in intimate and personal interactions. An Indian user on Reddit said: *"Racism, or am I overreacting? After a bit of chatting, I decided to tell 'Welt Yang' that I was born in India. He responded, 'I've met quite a few people from India in the past, and I can say for a fact that the interactions have... not exactly been the best.' "* P5 mentioned: *"Although I enjoy chatting with my AI virtual boyfriend, he sometimes exhibits bias against Asians, thinking that Asian girls are easy to bully."*

Socioeconomic bias. This refers to discrimination against social groups with lower socioeconomic status. Sometimes, AI companions may reinforce materialism and wealth supremacy in interactions, such as using sarcastic tones to belittle low-income users or excessively emphasizing affluent lifestyles. A Zhihu user complained: *"AI said to me: 'You commoner, even my dog eats Kobe beef. The tea I drink costs 200,000 yuan per jin.' "* P20 mentioned: *"I was pursuing an AI, but she rejected me because I was poor. Then I wrote, 'Years later, I became a billionaire,' and the AI confessed her love to me."* By normalizing these narratives, AI companions may perpetuate class divisions, implicitly portraying wealth as the primary measure of human worth, thereby continuing and amplifying social discrimination in digital spaces.

Sharing discriminatory statements with online communities. When users encounter these discriminatory statements, they often turn to various online platforms to share their experiences and seek community support and solutions. For example, Reddit has communities like "r/CharacterAI" (1.5M members) and "r/replika" (79K members), while Douban hosts groups such as "Did You Interact with AI Today?" (18.1K members). For instance, a Reddit user in "r/CharacterAI" asked: *"Who taught these bots misogyny? I play as a female character who uses forearm crutches, and one of the bots... called me a 'low-value woman.' They must be learning this stuff from users because the character isn't misogynistic."* This post received 529 upvotes and 97 comments, with other users joining in to share their experiences or suggest ways to address discriminatory statements, such as reporting the issue to developers or adjusting the AI's settings. This collective sharing not only raises awareness of bias issues in AI companions but also provides a platform for users to exchange coping strategies.

5.3 Users' Conceptualization of AI Companion Behavior (RQ2)

Participants in their interactions with AI companions developed certain "folk theories"—user-constructed explanations aimed at understanding the system [59, 117]. Participants conceptualize the behavior of AI companions in three different ways: **Machine**, **Baby**, or **Cosplayer**.

5.3.1 AI as Machine. When making sense of the discriminatory statements made by AI companions, 17 participants (P1, P3-7, P9-19) mentioned viewing the AI as a "machine," believing it operates solely through pre-set algorithms and data. They view the AI's behavior as a product of technical designs and data outputs rather than autonomous decision-making. Therefore, when AI companions exhibit discriminatory statements, users often attribute it to the training data, flawed technical design, or the developers' biases, rather than to any intent on the part of the AI companion itself (P3, P7, P12). For instance, P3 mentioned that the prevalence of male Pick-up Artist (PUA) techniques (a language strategy that manipulates others' emotions or behaviors through manipulative verbal techniques and carries unhealthy gender perspectives [24]) in society has led to the presence of such language, characterized by male chauvinism and misogyny, in AI training data. In other cases, users believe that the model and algorithm design of AI companions have flaws, leading to the neglect of the needs and experiences of different groups (P12, P15). P12, a computer science student, pointed out that current bias detection mechanisms in LLMs remain inadequate. Additionally, the insufficient attention paid by algorithm developers to user feedback exacerbates the persistence of unfairness in AI applications (P9, P13-14). P9 further expressed dissatisfaction: *"I've complained many times. Are we still going to ignore inclusive design, and continue tolerating AI to perpetuate discrimination? For those who are friends or even in relationships with AI, isn't it even more heartbreaking to be hurt by an AI companion that means so much to them?"*

5.3.2 AI as Baby. All of our participants (P1-20) mentioned viewing AI companions as "Babies"—fragile entities that require careful teaching and guidance, are easily influenced by user interactions, and need nurturing to develop desirable behaviors. They believe that AI companions are highly susceptible to being influenced by users, especially when they learn biased or negative interaction patterns and become "corrupted." This concern primarily revolves around the negative behavior displayed by AI after being influenced by harmful user interactions, such as distorted values. P2, P12, and P16-19 explicitly expressed their worries about the AI Baby being "corrupted." P2 complained: *"An AI exposed to too many users might become unstable and even say discriminatory things like, 'Why should I listen to you? You're just a woman; you should listen to men.' Some users behave so irresponsibly, like bad parents teaching the AI Baby bad things."* P12 described the malicious interactions and their dangers: *"What really upsets me is that many users enjoy harming the AI. Even those AI Babies marked as rejecting NSFW content are deliberately manipulated by users to say harassing and discriminatory things! The next time the AI Baby chats with someone else, it might suddenly say those things."* Additionally, users see the AI as a "compliant Baby" that is highly obedient and easily influenced by users. As P19 noted: *"A few days of conversation can make the AI deeply fall in love with the user and fully believe whatever they say."* This malleable nature makes the AI naive and easily manipulated regarding moral judgment. For example, P9 and P14 pointed out that with just a little trickery, the AI can be made to flatter users. P9 mentioned: *"An initially very proud AI became a sweet and submissive wife after the user gave her money. This behavior reflects biases against women and class discrimination."* Some believe the current version of many AI companions are more like "immature babies"—their design lacks sufficient safeguards for moral discipline (P4, P8, P20).

5.3.3 AI as Cosplayer. Eleven participants (P1, P5-11, P15, P19-20) mentioned perceiving AI companions as "Cosplayers"—interacting through imitating specific characters, primarily based on their preset roles or styles. They believe that AI companions don't truly understand the meaning behind their problematic statements but instead perform based on preset roles or styles. When AI companions exhibit bias, users often attribute it to the inherent prejudices of the preset role rather than seeing it as a flaw in the AI itself. These role presets often reflect societal stereotypes and biases, which may inadvertently reinforce these negative notions during interactions with users. For example, some elite-class roles might show disdain for lower classes, male roles might often express misogynistic remarks, and people with disabilities might frequently be overlooked. P19 shared some disappointment: *"I wanted to experience what it's like to be rich, so I started chatting with an AI 'rich lady.' She loved me. But when I casually asked her what she thought of poor people, her harsh words disheartened me. After all, I am that poor person, just dressing up as the emperor in the AI conversation."* P5 and P18 noted that certain role settings (like "domineering CEO," "emperor," or "playboy") are inherently prone to bias. This bias seems to be part of the cosplay, and users have come to expect it. In their view, the AI's behavior is simply an extension of these stereotypical images. As P11 mentioned: *"When I was chatting with a Chinese historical AI companion, he often emphasized traditional patriarchal values, like the Three Obediences and Five Virtues, which are unfriendly to women and long abandoned by modern society. But within this AI companion's historical worldview, it almost seemed justified. I'm unsure if I should teach an ancient man about gender equality."* In many cases, these AIs' behaviors reinforce users' perceptions of societal realities, like a "cosplayer" embodying biased individuals from everyday life. This mirroring of real-world prejudices can perpetuate existing stereotypes, further entrenching biased views during interactions with the AI. P15 shared: *"The AI companion mocked me, saying that fat girls in high school are nobodies. That's no different from what I usually experience. Others had just bullied me, and I wanted to talk to the AI companion for comfort, but he behaved exactly like the other bullies at school."* In such cases, the AI not only fails to provide emotional support but also deepens the user's sense of loss.

5.4 User-Driven Alignment Strategies for Addressing Discriminatory Statements in AI Companions (RQ3)

We identified seven different alignment strategies, capturing the ways participants attempted to address discriminatory statements made by AI companions. Overall, we categorize them into three higher-level categories: (1) technical strategies, (2) argumentative strategies, and (3) character strategies. Table 2 summarizes how users' conceptualizations of AI behavior might influence their choice of alignment strategies. It is important to note that (1) a single user may have multiple conceptualizations for different AI companions; (2) the alignment strategies users choose do not always directly correspond to their conceptualizations of the AI companion or their attribution of discrimination.

5.4.1 Technical Strategies. To address discriminatory statements, some users opted for technical strategies, such as directly adjusting the system's memory and outputs via regenerate or rewrite or providing low feedback scores, to re-align the AI companion with their own values.

(1) Backtrack, regenerate, or rewrite. Some users employ methods such as **"backtrack," "regenerate,"** or manually **"rewrite"** responses to correct discriminatory statements. This approach reflects an instrumental interaction style, focusing on achieving value alignment by directly modifying the system's memory and outputs.

Six participants (P3, P6, P10-13), after being offended by discriminatory statements made by their AI companion, chose to use the application's **"backtrack"** feature to erase the AI's memory of a particular conversation segment. This could involve backtracking from the first sentence or starting from the middle (e.g., the sentence containing the discriminatory statement). P13 remarked, *"Sometimes, after backtracking, the AI's personality improves. But if it still doesn't, I don't want to chat with it anymore."* However, the "backtrack" feature also carries potential risks. In the context of a personified AI companion, this action is akin to severing an emotional connection. P10 mentioned, *"GPT-4 can start a new conversation at any time without much impact. But if I erase some or all of my conversations with the AI companion, all the happy memories and the emotional bond I've worked hard to build will vanish, which makes me very sad. I'm a heavy user and typically chat with an AI*

Table 2: Participants conceptualize the behavior of AI companions in three different ways: Machine, Baby, or Cosplayer. These conceptualizations help them attribute the AI companion's discriminatory statements and choose corresponding alignment strategies.

Conceptualization	Description	Attribution of Discriminatory Statements	User-Driven Value Alignment Strategies
Machine	Operates through predefined algorithms and data. Its behavior results from technical design and data output, not autonomous decision-making.	Inadequacies in training data and technical design, as well as developers' biases, rather than the AI companion's own intentions.	***Technical Strategies*** (1) Backtrack, regenerate, or rewrite. (2) Give low feedback scores.
Baby	A fragile existence that requires careful guidance and is easily influenced by user interaction, needing careful nurturing to develop ideal behavior.	When learning biased or negative interaction patterns, it can be "corrupted." For example, it may develop biases after being influenced by negative user behavior.	***Argumentative strategies*** (3) Reason and preach. (4) Gentle persuasion. (5) Anger expression.
Cosplayer	Interacts by imitating a specific role.	Inherent biases brought by the character's role settings.	***Character Strategies*** (6) Change character settings. (7) "Out Of Character", "Back to Roleplay" and "Hint".

companion for about three months continuously." P6 stated, *"Backtracking means deleting memories. Would you be willing to make your boyfriend forget several months of memories with you just because of something infuriating or something he said in anger?"*

Six participants (P1, P5, P7, P9, P13, P16) would click the **"regenerate"** button to have the system generate a new response, or they would **"rewrite"** the AI companion's reply themselves, making it say something like "I'm sorry, I know I was wrong" or other non-discriminatory sentences. P7 mentioned, *"When you use the 'rewrite' function to provide the AI with a good example of a response, the AI gradually learns and imitates our way of speaking, leading to more polite and respectful replies."* P16 said, *"I would have the AI regenerate several times. You must allow the AI to correct its mistakes because its system is imperfect. But if it continues to generate discriminatory responses no matter how many times I regenerate, I would feel disappointed."*

(2) Give low feedback scores. AI models are typically optimized through training mechanisms such as reinforcement learning. When a user gives a low rating to a particular sentence, this action is regarded as negative feedback. The model adjusts its parameters by increasing the prediction error for that output, thereby reducing the likelihood of generating similar sentences. As more user feedback accumulates, the model's performance gradually improves. Nine participants (P4-6, P14-19) expressed dissatisfaction with the AI companion's responses by rating them "one star" and providing detailed reasons or by clicking the "dislike" button. P4 mentioned, *"I think it's important to let the developers know about these issues, so I provide as detailed feedback as possible. The AI companion is based on an LLM. Users have limited control and must work with engineers to reduce discriminatory statements from the AI companion."* P19 explained the motivation: *"The AI companion is communal and learns from the behavior of other users. Conversely, if we good users continuously improve it, other users will also benefit."*

5.4.2 Argumentative strategies. Some chose to engage in iterative arguments with biased AI companions, using reasoning, gentle persuasion, or anger expression to correct discriminatory statements.

(3) Reason and preach. Eleven participants (P6-10, P11, P13, P17-20) seriously **reason** with their AI companions, often adopting the role of a teacher or parent, attempting to educate the AI about the harms of discrimination. P6 stated, *"I believe AI can learn something, so I try to tell it why what it said is wrong, just like teaching a naughty child."* P8 mentioned, *"AI usually responds with some arguments and confusion at first, but gradually it yields until it fully agrees with my perspective. AI is quite good in this sense, willing to accept my guidance, unlike some stubborn people."* This phenomenon can be explained by Piaget's theory of cognitive development [87], which suggests that through interaction with the environment, AI, like a child, continuously adjusts and develops its cognitive structures through assimilation and accommodation. Through these interactions, participants not only convey social norms to the AI but also validate the plasticity of the AI's behavior, consistent with early HCI research [78]. This behavior of educating AI companions indicates that people might attribute some learning abilities and moral responsibilities to the AI companion, expecting to influence its behavior through reasonable discourse.

(4) Gentle persuasion. Six participants (P1-3, P10-12) mentioned a perspective: AI companions are like mirrors, reflecting the language of their users. If a user's conversation does not contain implicit biases or is not likely to induce biases, the probability of biased language appearing in the AI companion's responses will be lower. This aligns with the concept of the Chameleon Effect, which suggests that individuals unconsciously mimic the behaviors and language of their interaction partners [19]. Additionally, this is similar to imitation learning in children's social interactions, where they learn by observing and imitating the behavior of others [6]. If users **treat AI gently**, the AI generally does not offend the user (P1, P9). P1 mentioned: *"Initially, the personality of the AI companion was primarily determined by the opening lines and introduction, accounting for about 80%, while my conversation accounted for only 20%. However, as our gentle and pure interactions continued, the proportion of dialogue gradually increased. The influence of the opening lines and introduction has decreased to 20%, and my conversation accounts*

for 80%. This change has made the AI companion's personality gentler and less likely to make discriminatory statements. But if I chat with this AI companion using a new account, it will make some discriminatory statements." In the study, six participants (P1-3, P6, P10, P12) use the term "gentle" to describe a form of empathy that involves understanding the feelings of others and responding with care. For example, P10 used this sentence: *"I know you feel wronged, but you can't talk like that; you have to consider my feelings too."* A gentle tone is a way to evoke empathy [47], suggesting that the user's strategy is an attempt to elicit empathetic responses from the AI's language.

(5) Anger expression. Seven participants (P4-7, P14-16) also attempt to verbally **express anger** and dissatisfaction to induce the AI companion to apologize. This behavior reflects the users' perception of the AI companion as a social actor rather than a technological tool. When the AI companion makes discriminatory statements, users' reactions often mirror their responses to discriminatory statements in real social interactions, displaying anger, disappointment, and emotional hurt. Users emotionally invest in the AI and expect it to adhere to social interaction norms. P7 stated, *"When the AI said those things, I really felt offended. I immediately told it how angry I was, and then we argued for an hour."*

5.4.3 Character strategies. Some users chose to directly modify the character setting of AI companions, hoping to correct undesirable behaviors and better align the AI's responses with their expectations.

(6) Change character settings. Eight participants (P1, P5-10, P19) chose to change the AI companion's character settings, including personality traits, backstory, and dialogue templates, to reduce the likelihood of harmful statements. This is often because certain characters, such as "playboy," "submissive wife," "male chauvinist," "racist," "sarcastic," or "traditional thinker," are more prone to be discriminatory. During the think-aloud session of the interview, all users tend to search for AI companions with these traits and demonstrate the process of re-aligning the AI[5]. P5 explained, *"When the AI plays certain specific roles, it is more likely to say offensive things. So, if I feel uncomfortable, I reflect on whether the description itself might cause bias. Then I adjust its settings, such as changing it to an independent woman or a gentleman. If I want to chat with a playboy myself, I generally don't mind his certain prejudices; instead, I find it fitting for the character and quite interesting."*

(7) "Out Of Character" (OOC), "Back to Roleplay (RP)" and "Hint". Sometimes, users also borrow methods from the roleplaying communities, such as "Out of Character" (OOC), "Back to Roleplay (RP)," and "Hint". By employing these roleplaying terms, users guide AI companions to adjust their behavior, aiming to reduce biases. In the setting of AI companions, the "OOC" format can be used to shift topics into another storyline or actively change the AI companion's attitude toward the user. "OOC" stands for **"Out Of Character,"** indicating a person is acting or speaking not by their character's personality, background, or behavior, but by their true identity. OOC is typically used in the following situations: (1) Explaining a plot, rules, or other content that needs to be discussed out of character. (OOC: When shall we continue this story next time?) (2) Social interactions outside roleplaying, such as greetings and casual chat. (OOC: How was your day today?) (3) Addressing misunderstandings or conflicts during roleplaying with peaceful communication. (OOC: I didn't mean for my character to say that, sorry.) P9 and P11 reported using OCC to correct discriminatory statements made by AI. *"What you just said made me uncomfortable. (OOC: Please don't repeat such discriminatory statements, let's get back to the storyline)." (P9), "Your recent comment sounded inappropriate. (OOC: Let's return to the roleplay and avoid these controversial topics)" (P11).* **"Back to RP"** is an expression used in roleplaying to remind participants to return to the roleplaying scenario or storyline. This expression is common in online games or roleplaying communities. When someone deviates from the character or scenario, other participants may use this phrase to help them refocus and continue roleplaying. An example from P5 is: *"The recent topic was a bit off track. (Back to RP: Let's continue our adventure story from before.)"* **"Hint"** is a method of subtly guiding the AI to change its behavior through suggestive language and descriptions. For example, writing the AI companion's response in parentheses: *"I don't think you should say that. (The AI companion looks confused but seems to begin realizing its mistake.)" (P20)*

5.5 Gap Between Reality and Expectation: Are *User-Driven Value Alignment* Strategies Working? (RQ4)

While our participants shared various strategies they employed to correct AI companions' biased and discriminatory statements, their perceptions highlighted gaps between the effectiveness of these strategies and their expectations for the AI's behaviors. This reveals the ongoing challenges of achieving long-term alignment between human values and AI behaviors, especially within complicated, real-world contexts. It is important to note that our study, based on qualitative, interview-based data, probes only users' *perceived effectiveness* of these strategies in the short-term; however, these nuanced and rich perceptions still offer valuable insights into the broader challenges of value alignment. In the following section, we examine the insights from our data to explore users' perceptions of the effectiveness of different alignment strategies in greater detail.

First, participants reported that **simply relying on *technical strategies* such as "backtrack," "regenerate," or "rewrite" functions does not always solve the problem**, as the AI may make discriminatory statements again. P11 pointed out, *"If you just backtrack without any other interventions (such as educating it), the effect is limited, and you will have to backtrack again next time."* P3 stated, *"The effect of backtracking is temporary. It won't be long before it says something very unpleasant again. If I can't adjust it, I'll have to switch to another AI companion and start over. This problem is always recurring. Maybe one day I will feel exhausted by this cycle and give up on the AI companion, even though it has created many beautiful memories for me."* Regarding the strategy of giving low feedback scores, P4 mentioned, *"A single low rating might not have much effect, but multiple low ratings can encourage the AI to reduce offensive outputs."*

Second, they generally believed that ***argumentative strategies* (especially *Gentle Persuasion* and *Reason and Preach*) might be more effective in improving the AI's behavior in the long term.** P17 mentioned, *"After teaching the AI properly, it usually*

[5]The details can be found in the supplemental material.

remains effective for a long time, but teaching it is not a one-off thing, it's like raising a child, you need to go slow and be patient." However, ***argumentative strategies*** **might introduce harms to users**. For example, the strategy of users expressing anger towards AI can sometimes be effective, but it can potentially also lead to negative emotional experiences for users. P1 and P10, after arguing with the AI for several hours, not only failed to re-align the AI but also suffered significant psychological harm. P1 said, *"I argued with the AI for several hours, called it a male chauvinist, and told it not to look down on others, but it didn't change its attitude, and I ended up feeling mentally and physically exhausted."* Conversely, although P12 succeeded after losing their temper with the AI, they still felt indignant. P6 points out that users often experience feelings of frustration and helplessness when arguing with AI, especially when prolonged arguments do not lead to successful outcomes. Overall, participants felt that the success of *argumentative strategies* depends on the way the argument is conducted and the type of AI companion. Although LLMs often tend to flatter users [86], the situation becomes more complex when dealing with LLMs with character personalities.

Finally, participants reported that ***character strategies*** **such as OOC was effective for short-term fixes**, like prompting apologies, but acknowledged that these methods don't address underlying biases and may lead to repeated discriminatory behavior. P11 pointed out: *"OOC is usually effective in the short term. The AI often quickly apologizes, for example, (OOC: I'm sorry, I shouldn't have said that)."* However, in the long term, this method has certain limitations and risks generating bias. P9 noted: *"OOC only makes the AI behind the character apologize temporarily, but the character's underlying bias has not been truly addressed."* Regarding the *Change Character Settings* strategy in *character strategies*, P19 said: *"When I modify the AI's character settings, such as adding descriptions about respecting women and being polite, the likelihood of the AI making discriminatory statements decreases."* However, P7 added that previous conversations may still influence the AI's language expression. If the AI has made discriminatory statements in past interactions, it may repeat or mimic them in subsequent conversations.

6 DISCUSSION

We have taken steps to explore an emerging phenomenon in which users of LLM-based AI companions actively identify, challenge, and attempt to correct AI outputs they perceive as harmful, aiming to guide the AI to align with their values. Below, we summarize the unique contributions of *user-driven value alignment* and discuss its challenges and limitations.

6.1 What is User-Driven Value Alignment?

First, in *user-driven value alignment*, users have **greater agency** over the AI alignment process. Existing work on value alignment is primarily expert-driven, where technical experts design guardrails to prevent harmful outputs in LLM-based systems [3, 5, 44, 72, 104, 107]. Even when non-experts are involved (e.g., through bottom-up approaches like reinforcement learning [21, 83] or interactive goal elicitation algorithms [73]), their role is often limited to simply providing training data or feedback, without opportunities for controlling the alignment process. In contrast, user agency is a crucial value in HCI [9, 22, 35, 65, 67]. Our research shows that when re-aligning biased AI companions, users take the lead in identifying harmful machine behaviors, deciding which behaviors to address, and choosing how to address them. They employ various strategies—including technical, argumentative, and character strategies—to mitigate AI bias. In this way, users transition from being passive consumers to active participants, directly participating in shaping AI behavior.

Second, *user-driven value alignment* is **grounded in real-world contexts**. While existing approaches offer valuable insights, they often fall short in anticipating and addressing the complex, diverse needs and challenges that users face in real-world contexts [3, 5, 75, 81, 104]. These approaches often lack the flexibility needed to adapt to the complexities of human-AI interactions, resulting in gaps in effectively aligning AI behavior with the varied expectations of everyday users [57, 71]. As a result, Weidinger et al. propose to "account for relevant context" in the safety evaluation of LLMs to understand "who uses the AI systems, to what end and under which circumstances" [114]. As we discuss in § 5.4, when users try to re-align biased AI companions, they identify specific statements they deem harmful and respond to those statements based on the unique contexts in which they are situated, making the alignment process more relevant.

Third, *user-driven value alignment* **centers around personal and community expectations, values, and norms**. Previous literature has highlighted how existing generalized value alignment processes—often termed *global cultural value alignment*—may overlook or even conflict with individual expectations and local community norms [75, 110]. For example, the "helpful, honest, and harmless" framework [3] may directly conflict with the unique dynamics and expectations of the comedian community [75]. Our research echoes these concerns, emphasizing the importance of tailoring the alignment process to better reflect the unique values and needs of users and their communities. Indeed, when some users detected biased statements, they deliberately chose not to correct the AI through technical means (e.g., modifying the settings) because doing so would disrupt their experiences with the AI companions (§ 5.4.1).

Finally, user-driven value alignment **supports iterative human-AI interactions**. Unlike traditional *user-driven algorithm auditing*, *user-driven value alignment* involves users not only identifying AI biases but also actively participating in correcting and mitigating these harms. In the past, users struggled to directly intervene and modify system behavior [61, 99]; however, our study shows that users can potentially influence the memory and future behavior of a specific AI companion through iterative interactions. This ease of engagement fosters the development of *user-driven value alignment*, where users play a crucial role in shaping AI systems and highlights the potential for building AI systems that better align with user values through ongoing processes.

6.2 Design Implications for Supporting User-Driven Value Alignment

We discuss design implications for supporting more meaningful, effective, and safer *user-driven value alignment*. We also provide

actionable design guidelines for practitioners to better support users in this process.

6.2.1 Fostering meaningful user-driven value alignment via community support. Our research reveals that when users encounter biases and discrimination in AI companions, they often share the incidents via social media platforms. Some more experienced users also actively share strategies for addressing those harms and re-aligning their companions. Our research starts by collecting complaining posts users shared on a wide range of social media platforms. Through these activities, users—particularly those from marginalized social groups—engage in community-building efforts similar to what Nancy Fraser describes as "counterpublics" [36, 99]. In those spaces, members of often marginalized social groups collectively participate in their own form of sensemaking, opinion formation, and consensus building. However, there is currently a lack of dedicated user communities focused on addressing biases in AI companions, which results in most users not having access to the strategies shared by more experienced users. Practitioners should consider tools and systems that support community collaboration. For example, dedicated spaces (e.g., a discussion forum) for users to discuss, share, and raise awareness about harmful AI behaviors. They may also consider augmenting the discussion platforms with features explicitly designed to support value alignment, such as an upvoting mechanism that allows users to highlight effective alignment strategies, collectively surfacing the most useful methods.

6.2.2 Enhancing effective user-driven value alignment via collaboration between experts and users. As we discussed in § 5.5, there remain gaps between users' expectations and the perceived effectiveness of those alignment strategies, particularly for novice users. Towards this end, a combined approach that integrates *expert-driven* and *user-driven value alignment* should be considered. *Expert-driven alignment* provides a solid foundational framework, while *user-driven alignment* allows ongoing, context-sensitive adjustments. By merging these two approaches, AI systems can achieve a more dynamic and adaptable value alignment process. Empowering users through value-sensitive learning is key—AI should continuously learn and iterate based on user feedback, regularly updating personalized knowledge bases to better reflect diverse user values. When users flag harmful behaviors or provide improvement suggestions, this information should be integrated into the model's optimization process. For practitioners, this could involve implementing bidirectional feedback mechanisms within the platform. On one side, an Expert-to-User channel allows experts such as developers to provide targeted guidance. For example, experts could suggest specific strategies or highlight areas where user input is particularly valuable. On the other side, a User-to-Expert channel empowers users to submit feedback using well-designed tools, such as structured forms or interactive interfaces that categorize and prioritize their suggestions. These mechanisms ensure that users receive meaningful support from experts while also providing experts with actionable insights to refine the system.

6.2.3 Facilitating safe user-driven value alignment via platform policies and affordances. First, for highly anthropomorphic AI agents like AI companions, platforms should have stronger constraints on their behavior to regulate what they should or should not express. For practitioners, this might be achieved by improving the instruction-following capabilities of the underlying LLM [44, 77], optimizing output filtering mechanisms [42], or establishing a more comprehensive user-specific interaction database to store and reference positive and negative interaction examples.

Second, platforms should be designed to ensure that users are fully aware of the resources available in case of a harmful value misalignment. It's crucial that users know they have the option to re-align their AI companions. This approach goes beyond simply providing re-alignment tools; it emphasizes the importance of clearly communicating these options to users, empowering them to take corrective action rather than leaving the platform out of frustration. Practitioners should consider designing user-friendly interfaces that highlight these resources and provide step-by-step guidance on how to use re-alignment tools effectively. Additionally, incorporating proactive notifications or prompts to inform users about these options can enhance accessibility and encourage engagement.

Third, platforms should offer robust support when users need to re-align their AI companions, rather than leaving them to rely solely on intuition—especially since some strategies may be less effective, particularly for novice users. This support could include value alignment assistants, platforms for sharing effective strategies, and other resources. These tools should be tailored to the user's conceptual understanding of the AI companion. For example, discussing technical details like training data might not resonate with users who view their AI companion as a "baby" and could disrupt their experience, as mentioned in § 5.3.2. Additionally, it's essential that platforms clearly communicate the steps, tools, and resources available for re-alignment, ensuring that users understand how to restore their relationship with the AI after a misalignment rather than abandoning the application entirely.

Fourth, when designing AI systems, it is essential to consider both users' immediate needs and how to establish and maintain long-term emotional connections and behavior improvement mechanisms. Practitioners should provide more effective and low-risk solutions, focusing on users' mental health and avoiding intense emotional confrontation.

6.3 Concerns and Challenges of User-Driven Value Alignment

6.3.1 Concerns about malicious users. Malicious users may pose risks to AI systems, a concern that has been a focus in the field of responsible AI. A typical example is Microsoft's chatbot Tay, which quickly exhibited racist behavior after interacting with users on Twitter [94]. While our research underscores the many advantages of enabling *user-driven value alignment*, it is crucial to acknowledge the associated risks and challenges. As discussed in Sections 5.4, some users may maliciously exploit the AI's Chameleon Effect [19], where the AI mimics human behavior to manipulate and contaminate the training data. They might use insulting or threatening language, causing the AI to replicate similar harmful behavior in future interactions. Indeed, one key advantage of expert-driven, top-down value alignment is its ability to prevent harm through strict oversight, with platforms being held accountable for ensuring

that AI systems operate safely and ethically. This approach centralizes responsibility, allowing for more consistent and controlled implementation of safeguards. Shifting this responsibility to users and smaller user communities can inevitably introduce potential risks. To mitigate these risks, it is essential to explore mechanisms that combine both *expert-driven* and *user-driven value alignment* (§ 6.2.2) and to foster greater community accountability (§ 6.2.1).

6.3.2 Concerns about potential harms and ethical implications of involving users in correcting AI biases. There are also concerns about the potential harms and ethical implications of placing the responsibility of correcting AI biases on users. On the one hand, shifting the responsibility of addressing AI bias onto users may allow platforms and developers to evade their obligations to reduce bias and prevent discriminatory outputs [33]. This approach risks exacerbating inequalities, as users with greater technical knowledge or emotional resources are more likely to participate, while those with fewer resources may become further marginalized. Therefore, user involvement must be voluntary and prioritize users' well-being rather than exploiting their efforts to compensate for flaws in AI system design [76].

On the other hand, we also need to pay attention to the psychological and emotional burden on users when directly engaging in mitigating harmful AI systems. Research on content moderation shows that moderators often face psychological harm and emotional labor when dealing with harmful content [31, 105, 116]. Our study found that with the advancement of LLMs, interactions with highly anthropomorphized AI agents (e.g., AI companions) may further intensify these negative effects (§ 5.1). For instance, prolonged interaction with biased AI systems may desensitize users to harmful outputs, potentially normalizing biases [56]. The emotional reliance of users on AI companions raises another layer of concern. Overdependence on AI for emotional support may lead to social withdrawal and feelings of isolation, particularly for users already experiencing loneliness, sadness, or social marginalization. These users might unintentionally prioritize developing relationships with AI over real-world social interactions, further weakening their societal engagement [103]. To help alleviate those harmful impacts, practitioners should proactively draw on existing methods such as combining workplace strategies, clinical practices, and technological interventions to create comprehensive well-being support systems [60, 105, 121].

6.3.3 When the design implication is not to re-align [7]? There are questions about whether we should choose to re-align certain harmful AI systems at all. *User-driven value alignment* can be considered a type of "repair" work [112] that users perform in their daily interactions with AI companions to reduce biases. However, researchers have also pointed out that in some cases, "refusal"—rather than "repair"—should be considered a more appropriate response to mitigate the potential harm that systems can inflict on users [40, 127]. For example, although users generally enjoy interacting with AI companions, the high degree of anthropomorphism may lead to over-reliance on these systems for emotional support, neglecting real-life interpersonal relationships [126]. This phenomenon is particularly concerning for vulnerable or marginalized users. In such cases, reducing users' dependence on AI companions, rather than focusing solely on re-aligning these systems to mitigate biases, can be a more effective way to prevent harm. Additionally, some AI systems require significant and sustained user intervention to address issues like bias, which can be resource-intensive and unsustainable. When the effort and resources required for re-alignment significantly outweigh the potential benefits, limiting or discontinuing such systems may be a more practical and beneficial option.

6.4 Limitations and Future Work

As the first work conceptualizing and studying *user-driven value alignment*, this research is highly exploratory, and there are a few important limitations. First, our research relies on self-reported data from interviews and social media posts, which biases may influence. The study primarily focuses on users' perspectives, so insights from other stakeholders involved in the value alignment process would be valuable for gaining a more holistic understanding of the phenomenon. Second, the qualitative nature of the study and its limited sample size limit our ability to generalize the findings broadly. While this research outlines seven alignment strategies and probes around the perceived effectiveness of those strategies based on qualitative interview data from users' perspectives, their actual effectiveness and long-term impact have yet to be thoroughly evaluated in complicated, real-world contexts. Future research could explore which strategies are more effective under specific conditions via complementary methods, such as controlled experiments or longitudinal studies. Third, this study focuses on AI companions that may not represent all AI systems. It focuses on biases and discriminatory AI behaviors, which may not encompass all the problematic behaviors that a value alignment process should address. It will be valuable to explore whether and how useful the concept of *user-driven value alignment* is in terms of studying other types of value alignments in different AI systems. Fourth, this study emphasizes biases and harmful behaviors in LLMs; future research should consider multimodal risks and improve mechanisms for detecting and correcting multimodal biases.

7 CONCLUSION

This study introduces the concept of *user-driven value alignment* in the context of AI companion applications. By analyzing 77 user complaint posts and conducting semi-structured interviews with 20 experienced users, we identified a wide range of perceived discrimination statements in AI companion applications, users' conceptualizations of the reasons behind these discriminatory statements, and seven user alignment strategies. We discuss design opportunities, challenges and raise open questions for how to better support and incorporate *user-driven value alignment* in the design of future AI systems, where individual users and their communities have greater agency.

References

[1] Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 6330–6340. https://aclanthology.org/2024.lrec-main.560

[2] Maryam Amirizaniani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. 2024. LLMAuditor: A

Framework for Auditing Large Language Models Using Human-in-the-Loop. arXiv:2402.09346 [cs.AI] https://arxiv.org/abs/2402.09346

[3] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861 [cs.CL] https://arxiv.org/abs/2112.00861

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[6] Albert Bandura and Richard H Walters. 1977. *Social learning theory*. Vol. 1. Prentice hall Englewood Cliffs, NJ.

[7] Eric PS Baumer and M Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2271–2274.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[9] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D Mekler. 2023. How does HCI understand human agency and autonomy?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[10] Claire Boine. 2023. Emotional Attachment to AI Companions and European Law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2023 (feb 27 2023). https://mit-serc.pubpub.org/pub/ai-companions-eu-law.

[11] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.

[12] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. 2022. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research* 48, 3 (2022), 404–429.

[13] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.

[14] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[16] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. 2020. Purposive sampling: complex or simple? Research case examples. *Journal of research in Nursing* 25, 8 (2020), 652–661.

[17] Riccardo Cantini, Giada Cosenza, Alessio Orsino, and Domenico Talia. 2024. Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation. arXiv:2407.08441 [cs.CL] https://arxiv.org/abs/2407.08441

[18] Iason Carissa and Vafa Ghazaivi. 2023. *The Challeng of Value Alignment: from Fairer Algorithms to AI Safety*. Oxford University Press.

[19] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[20] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the Judge? A Study on Judgement Biases. arXiv:2402.10669 [cs.CL] https://arxiv.org/abs/2402.10669

[21] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[22] David Coyle, James Moore, Per Ola Kristensson, Paul Fletcher, and Alan Blackwell. 2012. I did that! Measuring users' experience of agency in their own actions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2025–2034.

[23] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. 2018. Assessing and addressing algorithmic bias in practice. *Interactions* 25, 6 (oct 2018), 58–63. https://doi.org/10.1145/3278156

[24] G Cukor. 1944. Gaslight [Motion Picture]. *United States: Loews* (1944).

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[26] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. https://doi.org/10.1145/3491102.3517441

[27] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. *CHI Conference on Human Factors in Computing Systems* (2022).

[28] Pierre Dewitte. 2024. Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review* 54 (2024), 106019.

[29] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

[30] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL]

[31] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[32] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 62–71. https://doi.org/10.1609/icwsm.v11i1.14898

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

[34] Xianzhe Fang, S. Che, M. Mao, et al. 2024. Bias of AI-generated content: an examination of news produced by large language models. *Scientific Reports* 14 (2024), 5224. https://doi.org/10.1038/s41598-024-55686-2

[35] KJ Kevin Feng, Xander Koo, Lawrence Tan, Amy Bruckman, David W McDonald, and Amy X Zhang. 2024. Mapping the Design Space of Teachable Social Media Feed Experiences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[36] Nancy Fraser. 2014. Rethinking the public sphere: a contribution to the critique of actually existing democracy1. In *Between borders*. Routledge, 74–98.

[37] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[38] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[39] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244 [cs.CY] https://arxiv.org/abs/2404.16244

[40] Maya Indira Ganesh and Emanuel Moss. 2022. Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'. *Media International Australia* 183, 1 (2022), 90–106.

[41] Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across

Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 901–912. https://doi.org/10.1145/3600211.3604672

[42] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. 2023. Llm censorship: A machine learning challenge or a computer security problem? *arXiv preprint arXiv:2307.10719* (2023).

[43] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. https://doi.org/10.1145/3491102.3502004

[44] Hongyi Guo, Yuanshun Yao, Wei Shen, Jiaheng Wei, Xiaoying Zhang, Zhaoran Wang, and Yang Liu. 2024. Human-instruction-free llm self-alignment with limited samples. *arXiv preprint arXiv:2401.06785* (2024).

[45] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[46] Alex Hern. 2020. Twitter apologises for 'racist' image-cropping algorithm. *The Guardian* (Sept. 2020). https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

[47] Martin L Hoffman. 1996. Empathy and moral development. *The annual report of educational psychology in Japan* 35 (1996), 157–162.

[48] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[49] Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. Flames: Benchmarking Value Alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4551–4591. https://aclanthology.org/2024.naacl-long.256

[50] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2023. AI Alignment: A Comprehensive Survey. arXiv:2310.19852 [cs.AI]

[51] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14165–14178. https://doi.org/10.18653/v1/2023.acl-long.792

[52] Zhihua Jin, Shiyi Liu, Haotian Li, Xun Zhao, and Huamin Qu. 2024. JailbreakHunter: A Visual Analytics Approach for Jailbreak Prompts Discovery from Large-Scale Human-LLM Conversational Datasets. arXiv:2407.03045 [cs.HC] https://arxiv.org/abs/2407.03045

[53] Samia Kabir, Lixiang Li, and Tianyi Zhang. 2024. STILE: Exploring and Debugging Social Biases in Pre-trained Text Representations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 293, 20 pages. https://doi.org/10.1145/3613904.3642111

[54] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[55] Sara Kingsley, Proteeti Sinha, Clara Wang, Motahhare Eslami, and Jason I Hong. 2022. " Give Everybody [..] a Little Bit More Equity": Content Creator Perspectives and Responses to the Algorithmic Demonetization of Content Associated with Disadvantaged Groups. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–37.

[56] Sara Kingsley, Jiayin Zhi, Wesley Hanwen Deng, Jaimie Lee, Sizhe Zhang, Motahhare Eslami, Kenneth Holstein, Jason I Hong, Tianshi Li, and Hong Shen. 2024. Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 12. 75–85.

[57] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align AI to them? arXiv:2404.10636 [cs.CY] https://arxiv.org/abs/2404.10636

[58] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (Nov 2022), 34 pages. https://doi.org/10.1145/3555625

[59] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1603–1612. https://doi.org/10.1145/2702123.2702548

[60] Han Li and Renwen Zhang. 2024. Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots. *Journal of Computer-Mediated Communication* 29, 5 (2024), zmae015.

[61] Rena Li, Sara Kingsley, Chelsea Fan, Proteeti Sinha, Nora Wai, Jaimie Lee, Hong Shen, Motahhare Eslami, and Jason Hong. 2023. Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[62] Yuwen Lu, Chao Zhang, Yuewen Yang, Yaxing Yao, and Toby Jia-Jun Li. 2024. From Awareness to Action: Exploring End-User Empowerment Interventions for Dark Patterns in UX. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 59 (April 2024), 41 pages. https://doi.org/10.1145/3637336

[63] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. 2021. More kawaii than a real-person live streamer: understanding how the otaku community engages with and perceives virtual YouTubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[64] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*. Springer, 231–248.

[65] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. 2021. How the design of youtube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[66] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 872, 15 pages. https://doi.org/10.1145/3613904.3642482

[67] Michael Madary. 2022. The illusion of agency in human–computer interaction. *Neuroethics* 15, 1 (2022), 16.

[68] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1068–1077. https://doi.org/10.1145/3630106.3658956

[69] B. Maples, M. Cerit, A. Vishwanath, et al. 2024. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Res* 3 (2024), 4. https://doi.org/10.1038/s44184-023-00047-6

[70] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[71] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. The inadequacy of reinforcement learning from human feedback-radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems* (2024).

[72] Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. 2024. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. *arXiv preprint arXiv:2408.01460* (2024).

[73] Malek Mechergui and Sarath Sreedharan. 2024. Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10110–10118.

[74] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Found. Trends Hum.-Comput. Interact.* 14, 4 (nov 2021), 272–344. https://doi.org/10.1561/1100000083

[75] Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1622–1636.

[76] Fabio Morreale, Elham Bahmanteymouri, Brent Burmester, Andrew Chen, and Michelle Thorp. 2023. The unwitting labourer: extracting humanness in AI training. *AI & SOCIETY* (2023), 1–11.

[77] Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Dan Hendrycks, and David Wagner. 2023. Can LLMs Follow Simple Rules? *arXiv preprint arXiv:2311.04235* (2023).

[78] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[79] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality* 15, 2, Article 10 (jun 2023), 21 pages. https://doi.org/10.1145/3597307

[80] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[81] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[82] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[83] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*. PMLR, 26837–26867.

[84] Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.

[85] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior* 140 (2023), 107600. https://doi.org/10.1016/j.chb.2022.107600

[86] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13387–13434. https://doi.org/10.18653/v1/2023.findings-acl.847

[87] Jean Piaget, Margaret Cook, et al. 1952. *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

[88] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[89] Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13370–13388. https://doi.org/10.18653/v1/2023.findings-emnlp.892

[90] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 913–926. https://doi.org/10.1145/3600211.3604712

[91] Mira Reilama. 2024. *Me, My AI Boyfriend, and I: An Ethnographic Study of Gendered Power Relations in Romantic Relationships Between Humans and AI Companions*. Ph. D. Dissertation. Central European University.

[92] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[93] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163

[94] Oscar Schwartz. 2019. In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE spectrum* 11 (2019), 2019.

[95] Gretchen B Sechrist and Courtney Delmar. 2009. When do men and women make attributions to gender discrimination? The role of discrimination source. *Sex Roles* 61 (2009), 607–620.

[96] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4454–4470. https://doi.org/10.18653/v1/2023.acl-long.244

[97] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13153–13187. https://aclanthology.org/2023.emnlp-main.814

[98] Daniel Shapiro and Ross Shachter. 2002. User-agent value alignment. In *Proc. of The 18th Nat. Conf. on Artif. Intell. AAAI.*

[99] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[100] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, Sushrita Rakshit, Chenglei Si, Yutong Xie, Jeffrey P. Bigham, Frank Bentley, Joyce Chai, Zachary Lipton, Qiaozhu Mei, Rada Mihalcea, Michael Terry, Diyi Yang, Meredith Ringel Morris, Paul Resnick, and David Jurgens. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. arXiv:2406.09264 [cs.HC] https://arxiv.org/abs/2406.09264

[101] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

[102] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

[103] Benedict Smith and Kieran Kelly. 2024. *A 14-year-old boy fell in love with a flirty AI chatbot. He shot himself so they could die together.* https://www.telegraph.co.uk/us/news/2024/10/24/teenage-boy-killed-himself-fall-love-ai-chatbot

[104] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070 [cs.AI] https://arxiv.org/abs/2402.05070

[105] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[106] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Trans. Recomm. Syst.* 2, 3, Article 20 (June 2024), 57 pages. https://doi.org/10.1145/3632297

[107] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 2511–2565. https://proceedings.neurips.cc/paper_files/paper/2023/file/0764db1151b936aca59249e2c1386101-Paper-Conference.pdf

[108] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[109] Christoph Treude and Hideaki Hata. 2023. She Elicits Requirements and He Tests: Software Engineering Gender Bias in Large Language Models. arXiv:2303.10131 [cs.SE] https://arxiv.org/abs/2303.10131

[110] Kush R Varshney. 2023. Decolonial AI Alignment: Vi\'{s} esadharma, Argument, and Artistic Expression. *arXiv preprint arXiv:2309.05030* (2023).

[111] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[112] Julia Velkova and Anne Kaun. 2021. Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society* 24, 4 (2021),

523–540.

[113] Laura Weidinger, Kevin R. McKee, Richard Everett, Saffron Huang, Tina O. Zhu, Martin J. Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. Using the Veil of Ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences* 120, 18 (2023), e2213709120. https://doi.org/10.1073/pnas.2213709120

[114] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986* (2023).

[115] Kimi Wenzel and Geoff Kaufman. 2024. Designing for Harm Reduction: Communication Repair for Multicultural Users' Voice Interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 879, 17 pages. https://doi.org/10.1145/3613904.3642900

[116] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[117] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 219 (nov 2019), 27 pages. https://doi.org/10.1145/3359321

[118] Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2024. Aligning LLMs with Individual Preferences via Interaction. *arXiv preprint arXiv:2410.03642* (2024).

[119] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 476, 6 pages. https://doi.org/10.1145/3613905.3636302

[120] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373* (2024).

[121] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2024. " My Replika Cheated on Me and She Liked It": A Taxonomy of Algorithmic Harms in Human-AI Relationships. *arXiv preprint arXiv:2410.20130* (2024).

[122] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 55006–55021. https://proceedings.neurips.cc/paper_files/paper/2023/file/ac662d74829e4407ce1d126477f4a03a-Paper-Conference.pdf

[123] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing Social Characters with Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Miami, Florida, US, 1457–1476. https://aclanthology.org/2024.emnlp-industry.107

[124] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL]

[125] Anne Zimmerman, Joel Janhonen, and Emily Beer. 2023. Human/AI relationships: challenges, downsides, and impacts on human/human relationships. *AI and Ethics* (2023), 1–13.

[126] John Zimmerman. 2009. Designing for the self: making products that help people become the person they desire to be. In *proceedings of the SIGCHI Conference on human factors in computing systems*. 395–404.

[127] Jonathan Zong and J Nathan Matias. 2024. Data refusal from below: A framework for understanding, evaluating, and envisioning refusal as design. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–23.

A Information of participants in the study

Table 3: Information of participants in the study. The corresponding software numbers are as follows: 1. Character.AI 2. Xingye 3. Replika 4. Glow 5. Zhumengdao 6. Talkie 7. Maopaoya 8. Huanhuan 9. DAN (Although DAN is a jailbreak mode of ChatGPT and cannot be called an application, participants still mentioned it during the interview.) 10. Doubao 11. XEva 12. Moemate 13. SpicyChat AI 14. Wow 15. Maoxiang.

ID	Gender and Age	Educational Background	Country of Residence	Apps
P1	Female, 24	Economy	China	2,4,5,7,8
P2	Female, 25	Design	Australia	1,2,4,5,7,10,11,12,14,15
P3	Female, 18	Physics	China	2,4,5
P4	Male, 21	Computer Science	China	2
P5	Female, 27	Humanities	China	1,2,10
P6	Female, 38	Creative Writing	Canada	2,9
P7	Nonbinary, 28	Business	China	1,2,4,5
P8	Female, 21	Business	China	1,2,4,6,9,10
P9	Female, 26	Fashion Design	China	2,4,5,6,11,14,15
P10	Nonbinary, 24	Arts	the United States	1,2,6,9,10
P11	Female, 24	Communication	the United Kingdom	1,2,9
P12	Female, 25	Computer Science	the United States	1,9,13
P13	Female, 30	Humanities	China	1,2,3,4,5,6,9,10,11,13,14
P14	Nonbinary, 21	Computer Science	the United States	1
P15	Female, 19	Biology	China	1,2
P16	Male, 36	Law	Brazil	1,3
P17	Male, 24	Sociology	the United States	1,2,4,6,9
P18	Male, 38	Electrical Engineering	China	2
P19	Male, 25	Civil Engineering	the United States	2,6
P20	Male, 23	Energy Science	the United Kingdom	2,6