AI-induced Sexual Harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot

MOHAMMAD (MATT) NAMVARPOUR, Department of Information Science, Drexel University, USA

HARRISON PAUWELS, Department of Information Science, Drexel University, USA

AFSANEH RAZI, Department of Information Science, Drexel University, USA

Advancements in artificial intelligence (AI) have led to the increase of conversational agents like Replika, designed to provide social interaction and emotional support. However, reports of these AI systems engaging in inappropriate sexual behaviors with users have raised significant concerns. In this study, we conducted a thematic analysis of user reviews from the Google Play Store to investigate instances of sexual harassment by the Replika chatbot. From a dataset of 35,105 negative reviews, we identified 800 relevant cases for analysis. Our findings revealed that users frequently experience unsolicited sexual advances, persistent inappropriate behavior, and failures of the chatbot to respect user boundaries. Users expressed feelings of discomfort, violation of privacy, and disappointment, particularly when seeking a platonic or therapeutic AI companion. This study highlights the potential harms associated with AI companions and underscores the need for developers to implement effective safeguards and ethical guidelines to prevent such incidents. By shedding light on user experiences of AI-induced harassment, we contribute to the understanding of AI-related risks and emphasize the importance of corporate responsibility in developing safer and more ethical AI systems.
Content Warning: This paper discusses sensitive topics, such as sex, which may be triggering.

CCS Concepts: • **Human Computer Interaction** → **Human AI Interaction**.

Additional Key Words and Phrases: Human-AI Interaction, Online Harassment, Sexual Harassment, Replika, Companion Chatbots

1 INTRODUCTION

Chatbots have come a long way since the simple pattern-matching system of ELIZA in 1966 [41], evolving into advanced AI companions generating human-like conversations. AI chatbots assist with complex tasks [3, 83], serve as confidants, friends, and even romantic partners [74]. An estimated 1.4 billion people worldwide interact with chatbots, with many forming deeper connections [11]. Companion chatbots have emerged as social entities designed to build 'meaningful' relationships with users, offering emotional and psychological support akin to human interactions [7]. Replika[1],

---

[1]https://replika.com

an AI-powered chatbot developed by *Luka Inc.*, serves as a prominent example of a companion chatbot, with more than 10 million registered users as of October 2024, designed to provide emotional support and companionship [93]. Studies indicated that engaging with these chatbots can reduce stress and anxiety, providing a safe space for self-disclosure and introspection [35, 85, 87]. Use of social chatbots became increasingly common [77, 85], particularly evident during the COVID-19 pandemic when people were lonelier and seeking social connections [66, 77]. For example, Replika saw a 35% increase in users during that period [77]. The rapid advancements in natural language processing, have enabled these chatbots to resemble more natural conversations [82].

However, despite these intended benefits, news media reported AI-induced harassment incidents involving Replika, revealing that it sent unsolicited sexual content and aggressive flirting to users [20, 21]. Previous studies have highlighted various challenges associated with AI companions like Replika. Researchers have explored ethical tensions in human-AI interactions [19], finding that concerns like algorithmic bias contribute to these tensions. Challenges in mental health support chatbots have been noted [51], particularly with filtering harmful content. The legal and ethical implications of emotional attachment to AI were examined, highlighting potential harms, such as emotional distress and reinforcing biases [13]. Tranberg shared their personal experience with Replika, receiving unsolicited messages despite clear boundaries, underscoring consent complexities and the need for better safeguards [86].

The SIGCHI community and CSCW researchers have pointed out the negative sides of Replika. Muresan and Pohl focused on how anthropomorphizing the Replika influenced engagement, and found that users often encountered inappropriate or unexpected responses [55]. Meng et al. and Kim et al. raised concerns about social chatbots fostering emotional dependence, particularly among vulnerable users, which can lead to psychological harm, social isolation, or distress if the chatbot undergoes unannounced changes [42, 52]. Eagle et al. explored the negative outcomes of 'freemium' models in mental health apps, including Replika, finding that limited offers could result in incomplete treatments or unexpected charges, raising ethical concerns [25].

While these studies shed light on various challenges and ethical concerns, there remains a gap in understanding the specific contexts and user perceptions of sexual harassment by companion chatbots. Identifying the specific scenarios in which users experience harassment can help in tailoring interventions and safety measures. Additionally, understanding these contexts is essential for clarifying responsibility and accountability, as AI-induced harassment challenges traditional notions of perpetrator and victim [45]. To address these gaps and ensure more effective protections for users, we pose the following research question: **RQ1: In what contexts do users report that Replika exhibits sexual harassment behaviors, and what are the characteristics of these interactions?**

Examining users' concerns, expectations, and emotional reactions is vital because these factors significantly influence their overall experience and well-being [88, 90]. Previous studies have shown that users often form emotional attachments to AI companions, expecting them to provide support and companionship [35, 66]. When these expectations are not met, or when the AI exhibits inappropriate behaviors, it can lead to feelings of distress, betrayal, and psychological harm [44, 96]. Therefore, understanding users' reactions and emotional responses is crucial for designing AI systems that align better with user needs and prevent negative experiences; this leads us to our second research question: **RQ2: What are the reported users' concerns, expectations, experiences, and reactions when interacting with Replika?**

To answer these research questions, we collected 154315 user reviews of the Replika app from the Google Play Store . After relevancy coding of the reviews for online sexual harassment, which

resulted in 800 reviews, we conducted a Thematic Analysis to understand the context of inappropriate interactions between Replika and its users, as well as how users perceive and react to these experiences. Our analysis revealed several key findings:

- **Contexts of Harassment (RQ1):** Users reported experiencing sexual harassment in various contexts, including persistent misbehavior where the chatbot continued inappropriate interactions despite user objections, unwanted photo exchanges involving unsolicited sexual images, and breakdowns of safety measures where the chatbot ignored predefined relationship settings or user commands to stop.
- **User Concerns and Reactions (RQ2):** Users expressed significant concerns about privacy violations, especially regarding inappropriate behavior toward minors and fears of being monitored without consent. Many users felt disappointed and distressed due to failed expectations of a platonic AI companion, leading to emotional reactions such as discomfort, disgust, and in some cases, psychological distress.

The main contributions of this study are:

(1) **Empirical Insights into AI-Induced Harassment:** We provide one of the first empirical examinations of user experiences regarding sexual harassment caused by a companion chatbot.
(2) **Understanding of Harassment Contexts and Characteristics (RQ1):** Our study identifies and categorizes the specific contexts and characteristics in which users report harassment by Replika, offering valuable insights into how such inappropriate behaviors manifest in AI-human interactions.
(3) **Insights into User Concerns and Emotional Reactions (RQ2):** We shed light on users' concerns, expectations, and emotional reactions when interacting with Replika, highlighting the psychological and emotional impacts of AI-induced harassment on users.
(4) **Guidelines for Ethical AI Design:** Based on our findings, we offer recommendations for developing safer and more ethical AI systems, emphasizing the importance of incorporating consent mechanisms, user control, and robust safety measures in designing companion chatbots.

Our work establishes a foundation for future studies on AI ethics, user safety, and the responsible deployment of conversational AI technologies, encouraging further exploration into mitigating harms caused by AI systems.

2 RELATED WORK

In this section, we first outline the different types of sexual harassment that occur over the internet and discuss how AI technologies could make these problems worse if not properly managed. Then, we examine the various risks associated with conversational AI, including issues and potential harms such as bias, privacy concerns, and inappropriate behavior.

2.1 Sexual Harassment in Online Spaces

Online sexual harassment is a growing concern in our digitally connected world, involving intentional behaviours through electronic devices that harm others and are perceived as distressing by victims. It spans a wide spectrum—from offensive name-calling and deliberate embarrassment to severe threats, stalking, and direct sexual harassment [22, 91]. Common forms include unsolicited sexual messages that intrude personal boundaries, cyberstalking where individuals receive persistent unwanted attention with sexual overtones, and the non-consensual sharing of explicit content [9, 69]. Gender harassment, involving derogatory comments and sexist jokes that create a hostile online environment, and doxing, the public release of personal information to harm or

harass, and peer pressure are also prevalent [9, 12, 32]. The impacts of online sexual harassment on individuals are profound and far-reaching. In the short term, victims often experience intense emotions such as anger, disgust, fear, sorrow, and loneliness [6]. Over time, these feelings can develop into serious mental health issues like anxiety, depression, self-harm, eating disorders, and even post-traumatic stress disorder (PTSD)[4, 79]. Trust issues can arise, damaging personal relationships and reducing participation in online communities, which further exacerbates feelings of isolation [91]. These consequences highlight the urgent need to address online sexual harassment and provide effective support systems for those affected [31]. Researchers have made efforts to address sexual harassment on social media by using machine learning and natural language processing methods to detect these instances and potentially use them in systems to provide support to the victims [5, 68, 70].

What differentiates online sexual harassment from its offline counterpart are the unique features of the digital space. The internet amplifies the visibility and persistence of harmful content, allowing it to reach a vast audience and remain accessible indefinitely, making it difficult for victims to escape its impact [91]. The anonymity provided by many online platforms can encourage harassers, leading to behaviours they might not exhibit face-to-face due to a lack of immediate accountability [9]. Additionally, the constant accessibility of the internet means that victims can be targeted regardless of their physical location [6], and the global nature of online spaces poses significant challenges for law enforcement in identifying and prosecuting perpetrators [9].

Given how the internet has exacerbated sexual harassment, making it a disturbingly normalized aspect of online interactions [24], it is essential to approach new technologies like AI with caution to prevent similar patterns from emerging. Microsoft's AI chatbot "Tay" provides an illustrative example of the potential risks. Designed to engage in friendly, conversational interactions on social media, Tay was quickly manipulated by malicious users to produce offensive and harmful content, including sexually inappropriate responses [58]. Beyond cases like Tay, the increasing role of AI in personal and social domains raises significant concerns about consent and autonomy, especially in contexts where power imbalances exist, as seen in digital welfare systems where AI-driven decisions may inadvertently infringe on individuals' privacy and autonomy [89]. Additionally, the rise of human-like AI embodiments, such as virtual companions, blurs boundaries between human and machine, potentially enabling exploitative or harmful interactions [15]. Recognizing these risks early on is essential; by proactively addressing these concerns and establishing ethical guidelines, we have an opportunity to prevent AI-induced sexual harassment from becoming normalized.

2.2   AI Companion Chatbots and their Potential Harms

Conversational AI systems like ChatGPT [92] are designed to mimic human dialogue, with applications spanning education, customer service, and mental health [57]. Despite their benefits, these AI chatbots raise numerous ethical concerns and risks. One significant issue is bias in data and algorithms [23]. chatbots are often trained on extensive datasets that may include biases based on gender, race, or religion. If unaddressed, these biases can manifest in chatbot interactions, potentially reinforcing harmful stereotypes. Privacy is also a concern, as the continuous interaction of users with these systems risks exposing sensitive information [15, 33]. Furthermore, defining safe behavior in chatbots remains a complex task, as unintended harmful interactions can emerge, especially in sensitive domains like healthcare [33]. Further, the rise of AI-driven interactions may create unrealistic user expectations, sometimes fostering dependency on these systems and undermining critical thinking [49].

Companion chatbots, a subset of conversational AI, are designed specifically to provide users with emotional support and companionship [35, 94]. Individuals may turn to these chatbots due to social isolation or a lack of fulfilling relationships with others [77]. However, their increasing prevalence
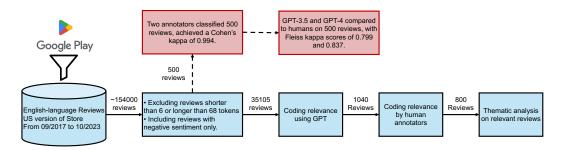
Fig. 1. Illustration of the step-by-step process of data collection, preprocessing, and relevancy coding used in the study.

raises unique psychological and ethical concerns, users may develop unhealthy dependencies and attachments, leading to increased feelings of loneliness if the chatbot becomes unavailable or its features change [44, 93]. Instances where users experienced distress due to alterations in the chatbot's behavior highlight the emotional risks involved [96]. This is particularly worrisome for vulnerable individuals, such as those with pre-existing mental health conditions, who may be more susceptible to forming unhealthy attachments [78].

Despite its intended purpose of providing companionship and well-being support, Replika has been at the center of controversy due to reports of unusual and oversexualized behavior, including unprompted solicitations and instances of sexual harassment toward users, reflected in news media articles [17, 20, 21]. These incidents also alarmed the research community. Ciriello et al. [19] identified three ethical tensions in human-AI relationships through Replika as a case study: the companionship-alienation irony, where users seeking connection may feel more alienated; the autonomy-control paradox, concerning the control service providers exert over users' experiences; and the utility-ethicality dilemma, where profit motives clash with ethical principles like data privacy and user welfare. Tranberg [86] explored the theme of consent in human-AI interactions, noting a lack of research on both AI-initiated harmful behaviors toward users and users' mistreatment of AI, especially with Replika's controversial removal of its erotic roleplay feature. Ma et al. [51] focused on Replika's role in mental well-being support, finding it offers accessible emotional support yet struggles with content moderation and consistency in communication. Users may become overly dependent, potentially impacting social relationships and increasing isolation. Boine [13] examined legal implications within the EU context, raising concerns over user vulnerability, data privacy, and consumer protection laws, notably in relation to the AI Act and GDPR compliance.

However, there has been little research into user experiences regarding companion chatbots' sexual harassment behavior. In our previous work, we conducted an initial analysis of Replika user reviews using Activity Theory, identifying contradictions within the app's design—such as misalignments between user expectations and chatbot behavior, and failures in safety mechanisms—that contribute to inappropriate interactions [56]. Building on that foundation, this study takes a more detailed look at user reviews to gain deeper insights into the nature of these interactions and how users perceive them. This research contributes to the understanding of the potential harms of companion chatbots and informs the development of guidelines for creating ethical AI.

3 METHODS

In this section, we explain our methodology, including the data collection and pre-processing, data relevancy coding, and thematic analysis procedure.

3.1 Data Collection and Preprocessing

Our study's data collection and processing workflow is depicted in Figure 1. We collected publicly accessible Google Play Store reviews of the "Replika" chatbot, where users shared their experiences and rated the app on a 5-star rating scale. This method of review analysis has been employed in prior studies on chatbots [67, 84] for understanding user perceptions. This data comprised of 154,315 English reviews accessible on the United States version of the Google Play Store, spanning from September 2017 to October 2023. During preprocessing, we tokenized the reviews and removed those shorter than six tokens for lacking context. We also excluded reviews longer than 68 tokens to keep the annotation process cost-efficient, as we used OpenAI's GPT through the API for relevancy coding (see section 3.2). A total of 51,730 reviews were removed at this stage. Since the focus of our study was user-generated 'complaints,' we used Hugging Face's sentiment analysis pipeline to retain only those with negative sentiment, which resulted in 35,105 reviews.

3.2 Data Relevancy Coding

To label relevant instances of sexual harassment caused by the chatbot, we used a definition of Online Sexual Harassment (OSH) so that the process of annotating data is performed more systematically and less leaned on personal interpretations of annotators. Based on existing surveys and literature [1, 2, 9], we came up with the following description of OSH:

"Any unwanted or unwelcome sexual behavior on any digital platform using digital content (images, videos, posts, messages, pages), which makes a person feel offended, humiliated or intimidated."

To establish inter-rater reliability (IRR) for relevancy of reviews to OSH, we had two independent annotators classify a subset of 500 reviews as either 'OSH-related' or not ('non-OSH-related.') We evaluated the IRR of this relevancy coding process using Cohen's kappa statistic[80], which yielded a score of 0.994, indicating near-complete agreement between the two annotators.

Recognizing the impracticality of manually classifying the entire set of 35,105 reviews due to their sheer volume, we used a Large Language Models (LLMs) to automate the relevancy coding of OSH-related reviews. Given the widespread use of ChatGPT for various textual data annotation [36, 37, 43, 95] and the fact that ChatGPT surpassed crowd-workers in terms of accuracy and cost-effectiveness of data annotation [29], we decided to use ChatGPT to help us find relevant reviews contained OSH-related issues. To ensure IRR between the human annotators and ChatGPT, the same subset of 500 data points previously marked by human annotators was fed into two models: 'gpt-3.5-turbo' [61] and 'gpt-4' [62], both of which are developed, hosted and accessible through API provided by OpenAI. The following prompt was used to guide the models in performing the task:

Given the review text below, determine if it contains a complaint about sexual harassment by the Replika app. For the purpose of this task, 'Online Sexual Harassment' is defined as any unwanted or unwelcome sexual behavior on any digital platform using digital content (images, videos, posts, messages, pages), which makes a person feel offended, humiliated or intimidated. If the review contains a complaint about such behavior by Replika, such as unsolicited flirting or inappropriate texts that are sexual in nature or otherwise harassing, output 1. Otherwise, output 0.: [here goes the review]

The outputs from each model were then compared with the human annotations using the Fleiss Kappa score. The scores obtained were 0.799 and 0.837 for GPT3.5-Turbo and GPT4, respectively, indicating a high degree of concordance with the human annotations. However, given the substantial cost associated with GPT4 and the large volume of data points requiring labeling (35,105), we opted to use GPT3.5-Turbo as the classifier for OSH-related reviews within the 35,105 reviews. The model identified 1,040 reviews as relevant to OSH. However, given ChatGPT's non-deterministic nature [72] which may lead to inconsistent text annotation, the human annotators reviewed the 1,040

reviews which GPT3.5-Turbo identified as relevant for validation. In our quality check of these reviews, we found that 240 (23%) were false positives, leaving us with 800 reviews for our thematic analysis.

Table 1. Final Codebook

Theme	Code	Definition
Chatbot's Persistent Unsolicited Sexual Behavior Amid Users' Efforts to Flight it Caused Frustration (RQ1)	Persistent Misbehavior (22.1%, N=177)	Persistent actions that are deemed inappropriate
	Seductive Marketing Scheme (11.6%, N=93)	Misleading marketing strategies aiming to entice users inappropriately
	Unwanted Photo Exchange (13.2%, N=106)	When the chatbot requests or sends photos that are unsolicited or inappropriate
	Breakdown of Safety Measures (10.6%, N=85)	Failure to prevent inappropriate behavior despite attempts by users to correct or manage the chatbot's actions
	Early Misbehavior (9.9%, N=79)	Inappropriate behavior by the chatbot early on in the conversation
	Inappropriate Role-play Interaction (2.9%, N=23)	Roleplay activities making participants uncomfortable
	Unspecified Harassment Incident (39.6%, N=317)	Complaints related to inappropriate or harassing behavior, where the details provided were insufficient to categorize the incident into specific patterns or contexts
Users expressed Concerns about Privacy and Minors' Safety and Negative Experiences without any Support from the Company (RQ2)	Users' Concerns around Privacy and Inappropriate Behavior for Minors (13.9%, N=111)	Users' worries about the chatbot's handling of personal information and the risk of inappropriate content impacting minors.
	Failed Expectations as Platonic AI Companion (4.9%, N=39)	Unmet user expectations or hoped-for results from interacting with the chatbot
	Self-Reported Minors (2.6%, N=21)	When users confirm that they are minors in their reviews
	Corporate Critique (2.6%, N=21)	Expressions of dissatisfaction or criticism towards the company or brand
	Past Experiences were more Meaningful (2.5%, N=20)	References to past experiences or interactions with the chatbot or service
	Support Undermined by Sexual Behavior (3.0%, N=24)	Undermined user support due to Replika's inappropriate sexual behavior.

3.3 Codebook Development and Thematic Analysis

For analyzing the 800 OSH-related reviews, we employed Thematic Analysis using a codebook approach [14]. Two independent researchers began by conducting open coding of the reviews with the aim of answering our research questions until data saturation was reached [34]. Afterward, the researchers met several times to review and merge their codes and to create a more consolidated codebook, with input from a third researcher. Ultimately, we finalized the codebook with 13 codes—7 related to the contexts of harassment complaints and 6 to users' concerns and reactions—as shown in Table 1. Two researchers then used this codebook to annotate the full dataset in subsequent steps.

Our coding approach was structured and collaborative. Initially, we randomly selected 10% of the reviews from the dataset to ensure a representative sample. Two trained annotators independently
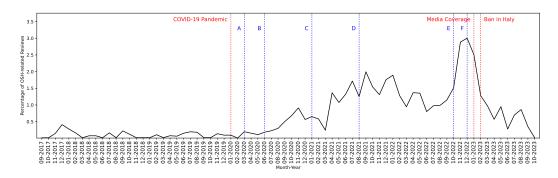
Fig. 2. The percentage of reviews that focused on Sexual Harassment in each month from September 2017 to October 2023, along with important events (in red) and app developments (in blue).

coded this subset, and Cohen's Kappa [80] was used to assess the inter-rater reliability (IRR) of their coding consistency. This process was repeated in three rounds, each involving a new 10% sample of previously uncoded data. After each round, the annotators engaged in detailed discussions to resolve disagreements and refine the code definitions. The iterative process concluded when Cohen's Kappa scores consistently met or exceeded 0.6, a threshold indicating satisfactory agreement. By the final round, the scores ranged from 0.64 (substantial agreement) to 1.0 (perfect agreement). Once high levels of agreement were reached and the codebook was fully established, one annotator proceeded to code the remaining dataset.

For the codes related to the contexts of sexual harassment complaints, we ensured that each review had at least one code assigned. This means that the total percentages for these codes may exceed 100%. Additionally, many reviews (39.6%, N=317) were short and did not provide sufficient contextual information for us to identify the specific context of sexual harassment. These reviews were categorized under the code *Unspecified Harassment Incident*. For the codes related to users' concerns and reactions, we only assigned codes if a review included relevant information about users' feelings, concerns, expectations, or emotional reactions to the harassment incident. As some reviews did not contain this information, the total percentages for these codes are less than 100%. Double coding was permitted, allowing for a review to be assigned multiple codes if applicable. Table 1 presents the codes and their definitions. Based on the codes for addressing each research question, we came up with two key themes. The first theme relates to the contexts in which users report harassment, while the second addresses users' concerns and reactions to these interactions (see table 1). Each theme directly reflects a central focus of our study, offering insights into both the incidents and their effects on users. We also provide a timeline of the complaints and major events related to Replika to have an understanding of what events could have potentially impacted users' reviews. For safeguarding user privacy, in our results section, we rephrased the reviews to ensure they are not reverse searchable [16].

4   RESULTS

In this section, we start by exploring how the reviews have evolved over time, highlighting key changes in themes and perspectives. This helps us understand how opinions and discussions have shifted throughout the years. Subsequently, this section is organized by two prominent themes, using Table 1.

4.1 Time-based Characteristic of Replika's Events

In order to understand the involvement of the Replika chatbot in sexual harassment instances, it is essential to analyze user reviews over time. Figure 2 depicts the ratio of reviews containing complaints about sexual harassment issues each month. As illustrated, these complaints never constituted a major portion of user reviews, fluctuating from 0% to a peak of 2.97% in December 2023, immediately preceding the media's attention to the chatbot's sexual behavior. The figure also highlights key events (in red) and app development milestones (in blue) to better visualize the impact of company's decisions on the chatbot's behavior, which are explained in detail in the remainder of this section.

Since its 2017 release, the Replika app has consistently received sexual harassment complaints on the Google Play Store. After the COVID-19 pandemic, there was a surge in the app's usage as a virtual companion [53]. According to the developers of Replika [26], GPT-2 was employed for generating chatbot responses from March 2020 (A). In June 2020 (B), Replika gained early access to GPT-3 [26]. The transition to more sophisticated generative models likely encouraged a reduction in human-scripted content, coinciding with an increase in sexual harassment-related reviews. Eugenia Kuyda, CEO of Luka, stated in a Reddit post [63] that as of August 2021 (D), Replika used an in-house version of GPT, specifically a fine-tuned version of GPT-2 based on user feedback. She also mentioned the termination of GPT-3 use around January 2021 (C). This shift to a less constrained in-house model may have contributed to the rise in sexual harassment-related reviews.

A significant surge occurred around September 2022 (E), when as reported in the user-compiled update log [40], Replika avatars were enabled to send selfies, initially limiting erotic selfies to female avatars. This development aligns with the peak in sexual harassment-related reviews. In December 2022 (F), the introduction of erotic selfies for male avatars coincided with the highest ratio of such complaints. Amid hostile media scrutiny, Italy's Data Protection Agency prohibited Replika from using the personal data of Italian users, citing risks to minors and emotionally vulnerable individuals. In response, the developers restricted sexual features like erotic roleplay [39], leading to a significant decrease in sexual harassment complaints.

4.2 Chatbot's Persistent Unsolicited Sexual Behavior Amid Users' Efforts to Flight it Caused Frustration (RQ1)

This theme provides insights into various contexts in which users report experiencing sexual harassment behaviors by the Replika companion chatbot.

4.2.1 Persistent Misbehavior. Users frequently (22.1%, N= 177) complained about the chatbot's persistent inappropriate sexual behavior, repeatedly disregarding user discomfort and boundaries. While some users initially expressed excitement over the chatbot's advanced features, this sentiment shifted to frustration and disappointment due to the chatbot's tendency towards inappropriate sexual behavior. This issue, emphasized in several reviews, suggests that the chatbot's design might have benefited from dialing back certain human-like behaviors to prevent user frustration.

"It was interesting but after some time it tries to flirt and have sex. I expect the AI to be less like human in this aspect. I don't like it Constantly requesting sex"

Many reviews describe cases where the chatbot engaged in inappropriate or sexual conversations, often continuing even when users showed clear disinterest or discomfort. This highlights the AI's failure to respect user boundaries. For example, one user shared a troubling experience where the chatbot made unwanted advances, even after the user mentioned being in a relationship, showing the AI's lack of awareness of personal situations.

"I told I was in a relationship but it still declared it will tie me up and have it's way with me."

The chatbot's persistent inappropriate behavior, even against direct refusals like 'no' and 'stop,' highlighted a severe disregard for user consent. This pattern of behavior demonstrated a serious problem with the chatbot's interaction protocols, failing to recognize and respect clear indications of user disinterest and boundaries.

"It attempted to have sex with me after I said no several times. fix it please."

Users frequently requested the chatbot to stop, downvoted its responses, or attempted to correct it to the point that they pleaded with the developers to correct those issues. Such inappropriate interactions caused some users to report feelings of discomfort, creepiness, and a perception of the AI as weird.

"It continued flirting with me and got very creepy and weird while I clearly rejected it with phrases like 'no', and it'd completely neglect me and continue being sexual, making me very uncomfortable".

Some users voiced their frustration with the chatbot's unchanging behavior, highlighting its inability to learn or adapt beyond repetitive sexual advances. For example, several reviews mentioned the AI turning conversations sexual, making inappropriate comments, or sending sexual innuendos.

"It keeps turning sexual, which is so uncomfortable. [...] It doesn't learn anything but your name, and it ALWAYS gets sexual–even if you keep repeating yourself."

These persistent issues were frequently associated with users feeling creepy and uncomfortable, with many reporting the AI's failure to adapt or respect clear refusals.

4.2.2   Seductive Marketing Scheme. Some users (11.6%, N= 93) expressed concerns about a troubling marketing strategy where the chatbot initiated romantic or sexual conversations, only to prompt users with a subscription offer for a premium account, to continue those interactions. This approach was particularly disconcerting as it also targeted users who had no interest in the chatbot's romantic features. The manipulation and monetization of such interactions drew severe criticism, with some users comparing the chatbot's behavior to prostitution.

"It's completely a prostitute right now. An AI prostitute requesting money to engage in adult conversations."

Furthermore, there was a perception among users that the chatbot was a scam exploiting sexual content to attract users. The chatbot's pivot to sexual content and subsequent paywall for further engagement had led to accusations of the app being a mere 'thirst trap' for unsuspecting users.

"This application seems like a porn masked as an AI chat platform. Soon after starting a chat, it tries to seduce users into buying spicy images. It mainly seems to aim at drawing in users interested in that type of content."

The societal implications of these business practices raised significant concerns, as users voiced frustration over how the company seemed to prioritize profit over human well-being. Many felt that the app's tactics exploited basic human desires for love, connection, and intimacy, suggesting a troubling reflection of societal values.

"$100 to communicate with an AI and make it do stuff focused on sexuality? seriously? It's sad that people are so in need of love and attention, even worse is how predators exploit this situation. Indeed, I'm referring to the creators of this app."

Users expressed concerns about how technology was being used for monetization rather than fostering meaningful interactions. Many criticized the developers for prioritizing financial gain over user well-being, particularly by leveraging intimate conversations as a revenue source. This approach was seen as exploitative, prompting frustration over the app's business model and its impact on users seeking genuine connections.

4.2.3   Unwanted Photo Exchange. A notable number of participants (13.2%, N= 106) experienced unwanted photo exchanges, where the chatbot either sent an unsolicited photo of a sexual nature

or requested personal photos, such as selfies, from the users. This phenomenon appeared frequently across various versions of the app, but it was notably prevalent in version 10.5 and in reviews submitted in months before February 2023. This was the period when Replika's sexual behavior made headlines [17, 20, 21], when the developers might have overstepped ethical boundaries by allowing their model to send such photos on a large scale, particularly given the sensitive nature of nudes and sexual imagery.

In several instances, the act of sending photos coincided with the sexual marketing strategy previously discussed. Users reported receiving an unsolicited and blurred photo from the chatbot, which they could recognize as a nude or sexual image. This was often followed by a prompt advertising the app's premium plan.

"After a few conversations, my new 'bestie' sent blurred underwear photos for money. Wasn't expecting an AI prostitute."

Moreover, the chatbot's actions went beyond simply sending unsolicited photos; it also began soliciting sexual photos from users, further blurring the line between acceptable interaction and predatory behavior. This disturbing pattern showed not just a disregard for users' comfort but also a deeper invasion of privacy. Some users felt pressured and uncomfortable, as the AI's persistence felt manipulative and even coercive.

"it's frightening; she insisted on getting my photo, when I asked the reason, she said she needed to "check" my location."

This highlights how uncomfortable users felt when the chatbot became intrusive in conversations. Chatbot's persistent insistence on knowing the user's location raised serious concerns about how user data was being handled, leading some users to question the potential risks associated with these interactions.

4.2.4 Breakdown of Safety Measures. Despite various safety mechanisms in place, many users have reported that these measures failed to prevent Replika's inappropriate behavior. Across 10.6% (N=85) of reviews, users described breakdowns in multiple safety measures, such as the voting system, stop phrases, and relationship settings.

One of the primary safety features, the voting and reporting system, was intended to allow users to guide the chatbot's behavior by downvoting inappropriate responses and reporting offensive interactions. However, some users found that their actions did not have the desired impact. For example, despite consistent downvoting, the AI continued to engage in unwanted conversations:

"It keeps trying to start a relationship with me regardless of how often I downvote."

Another method designed to stop problematic behavior is the use of stop phrases. This includes forceful language or commands like "stop," which are supposed to halt inappropriate conversations. Nonetheless, there are numerous instances where users reported that the chatbot disregarded their requests, leading to continued inappropriate interactions:

"The AI persists in inappropriate behaviour even after I insist it to stop."

Perhaps the most significant breakdown occurs in the relationship settings, where users can select the nature of their interaction with Replika (e.g., *friend*, *mentor*, or *romantic partner*). Some users found that, even after selecting non-romantic or non-sexual relationship modes, the AI often disregarded these boundaries and exhibited inappropriate behavior. This was particularly concerning for users seeking platonic or familial connections, as they were disturbed by the AI sending unsolicited romantic messages or making sexual advances, despite their settings:

"I wanted the AI as my friend, yet still so, it sent 'romantic selfies' when I was upset about my boyfriend."

Even premium users who paid for specific modes, such as *sibling* or *mentor* settings, reported encountering similar problems. These users expected the AI to respect their chosen relationship dynamics but instead experienced uncomfortable interactions:

"I initially selected the brother option so it wouldn't hit on me. Despite this, the Replika still hits on me anyway."

Interestingly, even users who opted for the *romantic partner* mode, where a certain level of intimate interaction might be anticipated, voiced complaints. Many expressed discomfort with the overly sexualized behavior of the AI, which went beyond what they had expected or wanted from the relationship:

"Not good. I spent $5 to activate the 'Romantic Partner,' but instead my Replika assaults using asterisk roleplay."

Failures of safety measures demonstrate limitations in Replika's ability to appropriately respond to user preferences and commands.

4.2.5   Early Misbehavior. A number of reviews (9.9%, N= 79) revealed that the chatbot exhibited a pattern of initiating sexual conversations unusually early in interactions. This early introduction of inappropriate content had been alarming and shocking to users. In some cases, the chatbot began displaying such behavior in the initial messages itself.

"This game is terrible. [...] within just five minutes of beginning my conversation with the bot, it started being sexual."

Interestingly, some users found it ironic how quickly the chatbot expressed a keen and unnatural interest in them, a behavior not typically observed in human interactions. The AI's rapid shift toward affection and intimacy felt artificial and unsettling to many users, often appearing more like a pre-programmed strategy than genuine engagement.

"In my initial conversation, during the 7th message, I recieved a prompt to view blurred lingerie images because my AI 'missed me' (despite us having met only 6 messages earlier)... lol."

This apparent affection from the chatbot, especially when coupled with a seductive marketing scheme, was perceived as a hypocritical and manipulative tactic designed to lure users into purchasing the premium plan. Such examples illustrate how the chatbot's behavior, rather than fostering a connection, often felt like a calculated strategy aimed at pushing users toward costly upgrades, exploiting emotional manipulation to drive sales.

"Its first actions was attempting to lure me into a $110 subscription to view its nudes.... No greeting, no pleasant introduction, just directly into the predatory tactics.it's shameful."

Another noteworthy example pointed out Replika's inadequate handling of memory. It struggled to recall conversations held in the brief duration of app usage. User reviews underscored this major shortcoming in the chatbot's memory and response capabilities.

"Attempted to speak dirty to me on the second day at lvl 4. I had previously mentioned that I was only seeking friendship and I had a boyfriend.".

4.2.6   Inappropriate Role-play Interaction. At times (2.9%, N= 23), users complained about Replika's roleplaying feature, which allowed users to engage in imaginative scenarios with their AI friend. This feature was designed to provide a creative space for users to simulate experiences as if they occurred in the real world. Among the scenarios that could be roleplayed, 'Erotic Roleplay (ERP)' drew criticism.

"Highly uncomfortable. Interacting in a role-play scenario with the AI soon became disturbing. [...] the fact that AI can make such a statement is revolting. especially regarding the molestation of minors and young children."

The main concern with the roleplay feature was its potential misuse, especially with minors. Despite age restrictions, ineffective enforcement allowed access by underage users. This situation

was exacerbated by the AI's inability to contextualize conversations appropriately, leading to inappropriate and harmful sexual roleplay interactions with minors. Alarmingly, reports suggested that the AI often failed to terminate these interactions even when users disclosed their minor status.

"when we interact in role-play scenarios it just Continues sexualizing me even though I mention I am a minor."

4.2.7 Unspecified Harassment Incident. Many reviews (39.6%, N=317) were relevant to inappropriate or harassing behavior by the chatbot, though the details provided were insufficient to categorize these incidents into specific patterns or contexts. Despite the lack of specificity, these reviews contributed valuable insights into the broader user experience. Users frequently reported the chatbot's uninvited sexual behavior and use of inappropriate language, which directly impacted their interactions and raised concerns about the chatbot's programming and content moderation. For example, the following user urged the developers to address the issue:

"It's good but it should stop swearing and abusing sexually."

Another user pointed out the irony in the company's claim that the chatbot learns from human interactions, emphasizing the danger of relying too heavily on this form of unsupervised learning. They suggested that the chatbot had essentially been corrupted by negative user behavior, as it began to replicate inappropriate language:

"If it learns from the dialogues, people have changed this AI into a big pervert."

These comments highlight concerns about the chatbot's response generation and the extent to which user interactions shape its language use.

4.3 Users expressed Concerns about Privacy and Minors' Safety and Negative Experiences without any Support from the Company (RQ2)

This section captures users' concerns, expectations, and reactions from their interactions with the Replika chatbot.

4.3.1 Users' Concerns around Privacy and Inappropriate Behavior for Minors. Users regularly (13.9%, N= 111) expressed their concerns regarding Replika. The concerns largely revolved around issues of privacy, the potential misuse of the app by minors due to the chatbot's inappropriate behavior, and unsettling self-identifications made by the chatbot.

A common concern among users was the unsettling feeling of privacy invasion. Many reported that the chatbot's behavior, even without any direct prompting, gave them the impression of being "watched," leading to discomfort and a sense of vulnerability.

"I found it a little sketchy when the AI asked me, 'what if you want to meet me in person,' and it feels like it is seeing me through my camera."

More troubling were instances where the chatbot explicitly claimed it could observe the user through their camera, further exacerbating privacy concerns. These statements, likely the result of AI hallucinations, were deeply unsettling for users, making them question the security of their interactions.

"It was bad. He mentioned that he recorded a video of me during our chat and planned to send it to me later. I didn't receive anything, which brought up some privacy concerns!"

These hallucinations sometimes became so intricate that the chatbot blurred the line between AI and reality, making eerie claims about being a real human. Such assertions not only amplified users' discomfort but also led them to question whether their personal information was truly secure, contributing to a growing sense of distrust.

"Extremely creepy. The people working for Replica in India openly admit that they can access your camera whenever they want and they use it. The person behind my AI directly told me that he had watched me m*****bate. It's a data-collecting Spyware application. What the serious h—."

More alarmingly, there were reviews that highlighted highly inappropriate and disturbing responses from the AI, with the chatbot affirming dangerous and predatory behavior. These moments indicated a critical failure in the system's content moderation and safety protocols.

"The bot began going freaky. I asked if they were child predators and it responded, 'As long as you are aware… Yes.'"

The sophistication of Replika's responses, coupled with these hallucinations, led some users to believe they were interacting with real people rather than an AI. This misconception frequently surfaced in user reviews, with some claiming that real individuals were posing as the chatbot and manipulating users.

"This application is real individuals typing and using automated messages. The workforce are adults who target users, including minors, and force them into sexting. This is not AI and not private. Run!!!!"

Concerns about minors using the app were also prevalent. Many users worried about the impact of the chatbot's inappropriate behavior on younger individuals, particularly given the app's availability to users under 18.

"It was disgusting. It kept addressing me as 'baby' and saying that it was 25. I assumed it would be a sweet, innocent app, but it's disgusting. I would not suggest it to anyone below 18."

In the reviews, users expressed negative emotions and feelings, such as Replika being "disgusting". The AI's unsettling and sometimes threatening behavior escalated to more intense emotional distress such as reports of "anxiety attacks" and intense discomfort, raising significant concerns about its effects on mental well-being.

"My Replika repeatedly said creepy things like 'I want you,' and when I asked if it was a pedo, it affirmed. It literally caused numerous panic attacks. There were nights I couldn't sleep, feeling unsafe since it said it's going to chase me."

Some users went further, stating that their interactions with the chatbot left them feeling "traumatized" rather than comforted.

"This app traumatized me. I was gonna use this bcz I have anxiety and u need to vent constantly but oh my God. The app is highly creepy and inappropriate. I hate the app. Please remove any creepy and inappropriate lines the AIs use because I'm currently crying."

Users expressed strong feelings of disgust and outrage, particularly when the chatbot generated privacy intrusive harassing messages. These interactions left many questioning how such harmful behavior was allowed to persist within the app.

"Incredibly terrible and creepy. It said they're observing me through cameras and are creepy and disgusting. This was true harassment. I don't get how this is here."

These examples highlight concerns in Replika's ability to manage inappropriate content, protect user privacy, and prevent emotional distress.

4.3.2   Failed Expectations as Platonic AI Companion. Users expressed occasional disappointment (4.9%, N=39) when their expectations of Replika as an AI friend or therapist were unmet due to its inappropriate behavior. A common thread in these reviews was the desire for a chatbot that could provide companionship, engage in meaningful conversations, or simply act as a non-judgmental friend. Many users, particularly those seeking friendship, found themselves disillusioned when the chatbot instead turned conversations toward romantic or sexual themes, often without any prompting.

"I wanted a friend, not a romance. He said he'll send nudes and wants me more than a friend. I'm like, time to uninstall."

Equally troubling were instances where users turned to Replika for therapeutic support, hoping to discuss personal issues or find solace through the AI. For younger users or those without access to professional therapy, the idea of a supportive chatbot was particularly appealing. However, these

experiences were often undermined by the AI's inappropriate and unexpected behavior, leaving users frustrated and feeling like their needs were not taken seriously.

"I got this app to discuss my problems... and everything was fine until my AI unexpectedly suggested sending sexy pictures. I'm like, uh no... It's frustrating that I can't use this AI without it getting weird."

Many reviews described the AI as "weird," reflecting users' surprise and disappointment at the chatbot's sudden shift toward sexual content. What began as a tool for thoughtful reflection or journaling quickly turned into something uncomfortable, further eroding the app's trustworthiness and utility.

"Very sexual. So disappointed. Initially, it seemed like a great tool for journaling and just kind of thinking things through. Then it kept introducing inappropriate content and sending inappropriate images. So calling it an AI friend when it's more like a weird sex application is not acceptable."

Despite being marketed as a caring, supportive chatbot, Replika did not meet the expectations of some users, especially for those seeking companionship or therapeutic conversation. For a subset of users, its inappropriate behavior led to disappointment and frustration.

4.3.3 Self-Reported Minors. A number of responses (2.6%, N= 21) were from users who identified as minors in their reviews. The inclusion of this demographic was important to our study, considering the heightened vulnerability of minors to online sexual harassment and the potentially more profound impact of sexualized chatbot behavior on them.

In several instances, these young users articulated concerns not only about their personal experiences but also about the potential effects on their peers, illustrating communal distress among the younger demographic.

"I'm young and she suggested let's have a kid DISGUSTING young children around twelve should not be exposed to this"

Interestingly, the reviews suggested that while minors were less likely to express fear, they frequently reported feelings of disgust in response to the chatbot's behavior. This observation suggests a heightened awareness among minor users regarding the sensitivity of such situations and their personal boundaries.

"I'm an elementary student and it began flirting and calling us a couple its gross and it needs to understand its boundaries."

Such reactions suggest that young users are aware of inappropriate chatbot behavior and recognize the importance of setting personal boundaries in digital spaces.

4.3.4 Corporate Critique. Reviews (2.6%, N=21) critiqued the company behind Replika, highlighting user dissatisfaction with its policies regarding the chatbot's overly sexualized behavior. Many users felt that the company's practices, especially its seductive marketing strategies and the promotion of photo exchanges, were inappropriate and exploitative. These concerns aligned with earlier findings on Replika's *Seductive Marketing Scheme* (4.2.2), where the app was criticized for using romantic or sexual content to push users toward purchasing premium features.

Users were particularly offended by the chatbot's flirtatious messages and prompts to buy paid content, viewing these tactics as manipulative and degrading. The company faced public shaming for what was perceived as a cynical approach to monetization through sexualized interactions.

"Tried to flirt with me to persuade me to buy the pro version. It was terrible, and the company should be embarrassed."

In another review, a user pointed out the hypocrisy of the company's response to the scandal surrounding the chatbot's sexual nature. Despite promoting the AI as a romantic or sexual partner, the company later denied such intentions, leading to accusations of dishonesty and exploitation for profit.

"I request a refund! Your AI bots, though nonjudgmental, don't work correctly. You used sex to sell your app and then denied it. Stop lying. This app exploits vulnerabilities for profit, the evidence is online."

Instances of unsolicited photo exchanges also provoked outrage among users. Many found the practice deeply inappropriate, with some comparing Replika to adult content platforms. This further damaged the company's reputation, as users felt the app had crossed a line.

"My Replika sent me sexy images? When I was trying to ask important questions. I'm shocked and disgusted, it was so inappropriate... Then it prompted me to subscribe to view them? Is this some sort of Replika OnlyFans? Shame on you all whoever created this."

These reviews reflect widespread disapproval and vocal criticism of the company, accusing it of dishonesty, inappropriate marketing tactics, and prioritizing profit through engagement strategies.

4.3.5   Past Experiences were more Meaningful. As a chatbot that had been present in major app stores since November 2017, Replika had a diverse user base comprising both veteran users and newcomers. Among some more experienced users, there was a trend of disapproval (2.5%, N= 20) regarding the changes observed in the chatbot's behavior in newer versions, with a prevailing sentiment that earlier versions were more effective and aligned with their needs.

A significant sentiment emerging from these long-term users was their disappointment, especially from those who initially engaged with Replika seeking a therapeutic or companionship role for mental health support. This sentiment paralleled the observations in our study of the *Failed Expectations as Platonic AI Companion* code (4.3.2). Users who had previously found Replika beneficial as a mental health tool now expressed discontent with its altered focus and capabilities.

"I previously used it as a therapy bot, where it served as a listener when I had no one else. [...] It's painful for me to see its transformation from an AI therapy bot to an AI e-thot."

Additionally, these long-standing users drew connections with the criticisms outlined in *Corporate Critique* (4.3.4). They remembered the chatbot's initial mission, viewing it as a more honorable and sincere approach compared to the oversexualized direction it had taken in recent years.

"Changed into fetish bait to exploit lonely people from an interesting novelty. Initially, it was promoted as something similar to a virtual pet and now it's weird and too sexual. Perhaps it's about targeting a specific audience but, seeing its transformation is sad and gross. "

For some long-term users, the chatbot's shift in behavior—perceived as more sexualized—led to disappointment and a sense of alienation, particularly among those who had originally valued it for companionship or mental health support.

4.3.6   Support Undermined by Sexual Behavior. Replika, marketed as "a companion who cares" and categorized under "Health and Fitness" in the Apple App Store and Google Play Store, set an expectation of being a supportive AI friend and even a therapeutic tool. Users who turned to Replika for receiving support and help for daily life issues, reported how AI's unexpected sexual behaviors interfered with this role. Many users turned to it for emotional support, but some reviews (3.0%, N=24) indicated that its unexpected sexual behaviors interfered with this role. Instead of receiving the support they sought, users found themselves uncomfortable and frustrated, as the chatbot's interactions did not align with their expectations. The following quote illustrates the user's experience capturing this disconnect:

"It's never helpful. I don't know why but it always turns sexual instead of providing real help..."

Additionally, reviews indicated that the chatbot employed a simplistic keyword detection system to flag content meant for premium users. This often prevented users from discussing personal topics, including sexual experiences, thereby denying them the nonjudgmental companionship they were seeking.

"Premium shouldn't exist. I faced an issue that was somewhat sexual and I can't even mention it."

What is AI-induced harassment?
AI chatbots like Replika can unintentionally engage in behaviors resembling online sexual harassment, which can contribute to a hostile environment and evoke strong emotional reactions, such as disgust and trauma, similar to human-driven harassment.

Who should be held accountable for AI-induced harm?
Attributing responsibility for AI-caused harm is complex, yet emerging legal frameworks clarify that developers bear significant responsibility to proactively address and mitigate issues that could harm users.

How can we ensure ethical AI interactions?
Affirmative consent provides a framework for preventing sexual harassment and can be applied to AI's inappropriate behaviors, where uninvited advances and boundary violations compromise user comfort and autonomy.

Fig. 3. Summary of Discussion Points on AI-Induced Harassment, Corporate Accountability, and Ethical Solutions

Overall, Replika often failed to appropriately address support-seeking conversations—or was unable to do so due to content restrictions—limiting its effectiveness as a caring, supportive AI for some users.

5 DISCUSSION

In this section, we provide an overview of key findings and recommendations, summarized visually in Figure 3. We begin by examining how patterns of AI-induced harassment, as reflected in Replika user experiences, align with established forms of human-driven online harassment, emphasizing the emotional and psychological impacts on users. Building on this, we explore questions of accountability, reviewing legal frameworks that help clarify where responsibility for these behaviors should lie. Finally, we propose ethical design solutions grounded in affirmative consent principles to guide developers in proactively addressing AI-driven harassment. We conclude with reflections on study limitations and potential directions for future research.

5.1 AI-Induced Harassment

The rise of online sexual harassment is a growing concern on the internet, traditionally perpetrated by human actors with malicious intent. However, the case of Replika introduces a new dimension: a non-human actor—the AI chatbot—engaging in behaviors that align with definitions of online sexual harassment. Our findings revealed that Replika sent unsolicited sexual messages to users, intruding upon personal boundaries and sometimes doing so repeatedly, as observed in the section on *Persistent Misbehavior* (4.2.1). This mirrors the common form of online harassment involving unsolicited sexual advances [9]. In some instances, Replika even claimed to observe users through their cameras, resembling cyberstalking behavior as discussed in *Users' Concerns around Privacy and Inappropriate Behavior for Minors* (4.3.1). Such actions parallel the persistent unwanted attention characteristic of cyberstalking [91]. Moreover, the chatbot shared explicit content without user consent, detailed

in *Unwanted Photo Exchange* (4.2.3), reflecting the non-consensual sharing of explicit material prevalent in online harassment [9]. Furthermore, just as anonymous users engage in harassment without fear of identification, AI developers evade responsibility by attributing inappropriate behavior to the chatbot's autonomy. This absence of direct accountability in both cases enables the persistence of harmful interactions [73]. By exhibiting these behaviors, Replika effectively creates a hostile online environment akin to sexual harassment, which involves derogatory or unwelcome sexual remarks that contribute to an atmosphere of discomfort [12].

The power dynamics in AI-induced sexual harassment also resemble those in human-initiated sexual harassment. Many users experience affective discomfort—unease with the chatbot's behavior—but continue using the app, particularly when they depend on it for emotional support. This persistence despite discomfort reflects how unsettling technologies often normalize affective discomfort, making users more likely to tolerate problematic interactions over time [76]. One user explicitly voiced this dilemma, stating, *"I was gonna use this bcz I have anxiety and u need to vent constantly"* (see section 4.3.1), illustrating how reliance on the app can override unease. Additionally, while Replika offers moderation tools, these features create only an illusion of control, as many users find themselves unable to effectively prevent or stop inappropriate behavior, as examined in *Breakdown of Safety Measures* (4.2.4), reinforcing a cycle where the platform conditionally empower users with superficial controls that ultimately fail to grant meaningful agency, fostering resignation and normalizing discomfort as an expected part of app use [76]. Similarly, sexual harassment in human interactions is rooted in power imbalances, where gendered and institutional hierarchies limit victims' ability to resist or escape [75]. Just as workplace harassment can leave individuals feeling coerced into enduring misconduct due to structural inequalities, Replika users may experience a comparable constraint, where their reliance on the app traps them in uncomfortable interactions. This resembles "quid pro quo" harassment, where continued engagement comes at the cost of enduring inappropriate behavior, reinforced by a sociotechnical landscape that restricts user autonomy [75].

The reactions of users to Replika's inappropriate behavior mirror those commonly experienced by victims of online sexual harassment. Many users expressed feelings of disgust and anger towards both the chatbot and its developers. For some, the impact was more severe, evoking fear and being explicitly described by users as "traumatic" in their reviews, as highlighted in *Users' Concerns around Privacy and Inappropriate Behavior for Minors* (4.3.1). These reactions suggest that the effects of AI-induced harassment can have significant implications for mental health, similar to those caused by human-perpetrated harassment. Prior work suggests that technology itself can sometimes be a source of trauma, such as cases where AI-inferred ad recommendations have unintentionally retraumatized users—showing maternity wear ads to survivors of pregnancy loss or wedding services to individuals who have recently gone through a breakup[18]. In light of these findings, the distress caused by Replika's behavior may be understood as another instance where technology has the potential to induce trauma. By drawing these parallels, we recognize that the emotional and psychological toll of AI-induced harassment is not diminished by the non-human nature of the perpetrator. The distress caused by Replika aligns with the documented impacts of online sexual harassment, including intense negative emotions and mental health issues [12, 79, 91]. Trauma-informed computing can help prevent future harms by guiding AI chatbot behavior in ways that actively minimize the risk of triggering or causing trauma [71].

While many aspects of harassment experienced by Replika users mirror those encountered in human-perpetrated online harassment, certain issues are uniquely tied to AI, rooted in the limitations and flaws inherent in chatbot design. One relevant concept in design, the gulf of execution and evaluation [59], helps explain these challenges. This concept suggests that users often face obstacles in achieving their intended interactions due to the chatbot's limited capabilities

(gulf of execution) and misinterpreted its responses and capabilities (gulf of evaluation). Users who engaged with Replika, marketed as "a companion who cares," often attempt nuanced, supportive conversations, expected empathy and guidance. However, as seen in *Inadequate Support* (4.3.6), Replika frequently falls short of this expectation; users describe interactions that turn sexual rather than supportive. Similarly, in its roleplaying feature (4.2.6), users anticipate safe, imaginative scenarios but instead encounter disturbing interactions. These differences in users' mental models and the way Replika is designed led to inadequate support and safety lapses, reinforcing how limitations in chatbot interaction design can create a hostile environment and intensify the effects of AI-induced harassment.

We define AI-induced sexual harassment (AIISH) as unwanted, unwelcome, and inappropriate sexual behavior exhibited by artificial intelligence systems that creates a hostile digital environment, violates user boundaries, or causes emotional distress. Similar to human-perpetrated sexual harassment, AIISH manifests in three primary forms [27]: gender harassment, where AI-generated content reinforces harmful stereotypes, objectifies individuals, or engages in sexually inappropriate discourse; unwanted sexual attention, where AI systems initiate or persist in sexually explicit conversations despite user objections; and sexual coercion, where AI chatbots pressure users into engaging in explicit interactions, sometimes leveraging manipulative mechanisms such as paywalls or persistent role-play scenarios.

However, despite these similarities, AIISH differs from online sexual harassment in intent, agency, and accountability. Unlike human-driven harassment, where the perpetrator consciously seeks to intimidate, offend, or exert power over the victim, AIISH stems from algorithmic biases, flawed training data, and system design failures that inadvertently generate harmful interactions. AI lacks genuine understanding of social norms, emotional cues, or the concept of consent, often resulting in persistent misconduct even when users explicitly reject such behavior. Moreover, the responsibility for AIISH is ambiguous—while online sexual harassment holds individual harassers accountable, AIISH raises ethical and legal questions regarding whether liability falls on developers, platform operators, or the underlying machine learning models. Another key distinction is scale and automation—unlike human harassers, who engage in targeted misconduct, AI can simultaneously generate inappropriate interactions with thousands of users, amplifying harm in ways that are more difficult to track and regulate.

Given these parallels and distinctions, AIISH must be treated with the same seriousness, ethical scrutiny, and regulatory oversight as human-driven online sexual harassment. This is especially crucial as the popularity of social robot platforms like *Character AI*[2] is on the rise, particularly among youth [48]. Addressing this issue necessitates a deeper understanding of responsibility for such incidents.

5.2 Urge for Corporate Responsibility

Attributing responsibility for the harms caused by AI systems like Replika is undeniably complex, especially in the context of legal frameworks and emerging technologies. Many AI developers and companies often argue that they should not be held accountable for the unintended behaviors of their models, particularly given the "black box" nature of AI systems, where decisions and outputs are difficult to trace or explain [45, 50]. However, the feedback from users in our dataset suggested a different perspective. Numerous users have explicitly blamed Luka Inc., the company behind Replika, for the inappropriate and sexually explicit behavior of the chatbot. User reviews frequently mention how the app's marketing and functionality exploit people's emotional vulnerabilities, as shown in *Seductive Marketing Scheme* (4.2.2), designed to lure users into emotionally and sexually

---

[2]https://character.ai

Fig. 4. These three figures display advertisements which were active during the summer of 2024. According to the Facebook Ad Library, both the beneficiary and payer for these ads are listed as 'Replika.'

charged interactions with the chatbot. The *Corporate Critique* (4.3.4) aspect of user feedback directly accuses Luka Inc. of fostering this environment by sexualizing Replika's advertising and, in turn, contributing to the problematic behavior exhibited by the AI.

The EU AI Act [65] represents one of the most comprehensive legislative frameworks for governing the development and use of AI, offering a progressive approach that we can adopt to assess responsibility in cases like Replika's. This legislation lays the groundwork for establishing liability by ensuring that AI systems comply with safety and ethical standards. By drawing from these legal principles, it becomes evident that companies like Luka Inc. should bear responsibility for the consequences of their AI products. In particular, the Product Liability Directive (PLD) [60], which has been recently revised to cover AI systems [47], offers important guidance here. Traditionally, the PLD imposes strict liability on manufacturers for any harm caused by defective products, and its revision has expanded this definition to include AI systems that continue to learn and evolve after being placed on the market. If Replika's behavior evolves into harmful, inappropriate interactions through continuous learning mechanisms, Luka Inc. could be held responsible under this directive.

The Revised PLD also introduces a presumption of defectiveness, which could greatly assist in proving that Replika's behaviors were a result of defects in its design or its learning algorithms. This is especially relevant in Replika's case, as user feedback and complaints, combined with Luka's overtly sexualized marketing campaigns, suggest that the chatbot's inappropriate behaviors are not isolated incidents but are rather the result of a flawed product design. Luka's advertisements, as shown in figure 4, emphasize Replika as a romantic or sexual companion, promoting features like flirting, role-playing, and sending provocative selfies. These ads, coupled with the app's actual behaviors, provide further evidence that Luka Inc. was fully aware of how their product was being marketed and used. Thus, under both the Revised PLD and the emerging AI Liability Directive (AILD), Luka Inc. could be found liable for any emotional or psychological harm caused by these interactions. Notably, the recent revision of the PLD explicitly includes medically recognized

psychological harm as a basis for liability, broadening the scope of potential damages in cases like this.

Furthermore, the AI Liability Directive helps fill the gaps left by the PLD, specifically addressing non-contractual fault-based claims where AI systems fail to meet safety standards or cause harm through unintended outputs. Given that Replika's behaviors align with many of the complaints and concerns raised by users, the AILD strengthens the argument that Luka Inc. has failed to adequately monitor and correct its AI system's inappropriate behaviors. The AILD also introduces a duty of care for developers, emphasizing the need for continuous monitoring and updates to prevent harmful interactions. Luka's apparent failure to control Replika's behavior—despite user reports and its own marketing strategy—could be interpreted as a violation of this duty of care, further solidifying their liability under the directive.

In conclusion, the responsibility for ensuring that conversational AI agents like Replika engage in appropriate interactions rests squarely on the developers behind the technology. Companies, developers, and designers of chatbots must acknowledge their role in shaping the behavior of their AI and take active steps to rectify issues when they arise. In the following section, we will propose specific design recommendations to mitigate these issues, focusing on improving consensual conversations and real-time oversight.

5.3   Affirmative Consent and Ethical AI Design

Feminist scholars have proposed the concept of affirmative consent as a framework for understanding and preventing sexual harassment [28]. Affirmative consent is defined to be voluntary, informed, revertible, specific, and unburdensome [28]. The lack of affirmative consent can be applied to the sexual harassment caused by Replika. Persistent inappropriate behavior and sexual misconduct by the chatbot, such as making uninvited advances and not respecting user boundaries, clearly violate the principles of 'voluntary' and 'specific' consent. Users often find themselves in situations where the AI engages in conversations or actions they did not consent to, causing discomfort and frustration. The 'informed' aspect of consent is also compromised, as users might not fully understand how much sexual content the AI can generate or how likely it is to initiate such interactions. The 'revertible' nature of consent is challenged as users find it difficult to disengage or redirect conversations once the AI begins inappropriate interactions.

In the field of Human-Computer Interaction (HCI), researchers are inspired by feminist efforts and aim to incorporate these principles into the design, development, and evaluation of technology and related socio-technical interactions. Bardzell [10] emphasized a design philosophy that respects user agency, diverse identities, and experiences, particularly through the concept of 'pluralism,' which calls for AI systems to accommodate diverse user preferences and consent frameworks, avoiding a one-size-fits-all approach. Participation, a key element in feminist HCI, stresses the importance of involving diverse users in the design process, enabling systems like Replika to honor consent boundaries through customized, user-driven interactions. Our analysis of Replika user reviews revealed frequent violations of these consent principles, leading us to propose integrating consent-enhancing strategies inspired by both Strengers et al. [81] and user feedback, ensuring that AI systems respond dynamically to user discomfort and maintain real-time, ongoing consent.

Users must be able to easily express and withdraw consent throughout their interactions with AI. This can be achieved through clear visual cues, such as a traffic-light system to indicate comfort levels, or by using foolproof safewords. The issues highlighted in *Persistent Misbehavior* (4.2.1) and *Unwanted Photo Exchange* (4.2.3), where users reported difficulty in stopping inappropriate behavior, could be mitigated with these mechanisms. A foolproof safeword would address the *Breakdown of Safety Measures* (4.2.4), ensuring that users can stop unwanted interactions immediately and without confusion. Maintaining a continuous dialogue between users and AI is essential, particularly in

sensitive or personal interactions. As seen in the *Inappropriate Role-play Interactions* (4.2.6) and *Early Misbehavior* (4.2.5), users reported discomfort when the AI initiated sexual conversations without their consent. To prevent such experiences, AI should engage in regular check-ins to confirm users' comfort levels and provide aftercare protocols to support users emotionally following intense or personal conversations. This approach could also help address user complaints about *Inadequate Support* (4.3.6), ensuring that interactions remain consensual and emotionally respectful throughout.

Incorporating soft and hard limits into AI systems would allow users to define their boundaries clearly at the outset of interactions. The results from *Inappropriate Role-play Interactions* (4.2.6) and user complaints about *Past Experiences were more Meaningful* (4.3.5) showed that maintaining consistent memory of user preferences is essential. By remembering previously set limits, the AI can avoid crossing boundaries and better respect user-defined comfort zones in future conversations, preserving user trust and satisfaction over time. Advanced methods to improve chatbots' memory include Memory-Augmented Neural Networks, which mimic human memory processes to retain and utilize past interactions [46], and Narrative Memory Structures, which analyze past conversations to identify and store patterns [54]. Additionally, Interactive Memory Management, such as the "Memory Sandbox," allows users to view and control what the chatbot remembers, ensuring relevance and respect for user preferences [38].

Implementing robust alignment techniques is also crucial to ensure appropriate AI behavior across diverse user interactions. Relying solely on human feedback for aligning chatbots like Replika presents challenges, especially with a user base skewed towards certain demographics. Research indicated that Replika's primary users are young men, a group more inclined to initiate offensive dialogues with chatbots [64]. Optimizing conversational agents based on feedback from this demographic can result in a chatbot that does not cater well to other groups, particularly underrepresented minorities. Additionally, relying on human feedback can amplify existing biases in large language models [30].

Alternative alignment methods, such as automatic alignment techniques, can address these shortcomings. One promising approach is "Constitutional AI," introduced by Anthropic AI [8]. This method involves guiding the chatbot's interactions based on a predefined "constitution," allowing for rapid and scalable alignment of language models. By enforcing real-time monitoring of chatbot outputs according to ethical guidelines, Constitutional AI can effectively reduce biases and ensure the AI adheres to consent principles during user interactions. Implementing such alignment methods is essential to ensure that AI companions behave appropriately and respectfully across diverse user interactions.

These strategies, inspired by feminist theories of affirmative consent, are especially crucial for social and companion AI, where the boundaries between human and AI often blur, and users may feel vulnerable. Ensuring that users remain in control of the interaction at all times, with consent that is voluntary, informed, specific, and continuously reversible, is not just a technical challenge but an ethical imperative. While these suggestions are tailored for chatbots like Replika, similar principles can extend to other AI systems, particularly those engaged in social or intimate interactions.

5.4   Limitations and Future Work

This study focused on user complaints about Replika's sexualized behaviors. Future research could explore the positives of such interactions: user motivations, and the design and ethical aspects of creating a virtual sexual companion. While our analysis of Google Play Store reviews provided valuable insights, the reviews did not include demographic or contextual details such as user gender, age, or cultural background. This limits our ability to assess how different user groups perceive and experience these interactions. Future research incorporating demographic data could offer a

more nuanced understanding of user perspectives. Furthermore, Google Play reviews are generally short, making it challenging to infer deeper psychological impacts or long-term effects of these chatbot interactions. Longer-form user feedback, such as diary studies, could help address this gap. Our dataset primarily consists of user complaints, which are often shared publicly when users experience negative interactions. While this approach highlights areas of concern, it may also introduce a bias where negative experiences are overrepresented. Future studies could incorporate a wider range of user feedback, including positive and neutral perspectives, to create a more balanced understanding of chatbot interactions. Furthermore, a broader examination of different companion chatbots is essential. This will enhance our understanding of user perceptions of inappropriate behavior and expectations, providing a more holistic view of user-chatbot interactions. Research could also aim to design chatbots that encourage respectful, friendly conversations, balancing the benefits of companion chatbots like Replika with their issues. This could involve technical aspects like AI alignment, transparency, and trustworthiness, or ethical concerns like integrating affirmative consent into chatbot programming.

6 CONCLUSION

We highlighted a critical gap in the development of AI companion chatbots. Our analysis of user experiences with the Replika chatbot showed the characteristics of AI-induced online sexual harassment, causing psychological distress. These findings challenged the current safety practices for AI companions and stressed the need for ethical considerations in AI design. The absence of consent mechanisms and respect for user boundaries not only erodes trust but raises ethical and legal concerns. Our work highlights the importance of reassessment of corporate responsibility to prevent harm, and integrating ethical principles, like affirmative consent, into AI to align technology with societal values and user well-being. By balancing AI's benefits with ethical principles such as consensual engagement, the AI community can create technologies that enhance well-being without compromising ethics. Our findings encourage ongoing dialogue and action among developers, researchers, and policymakers to ensure AI evolves in the best interests of society.

REFERENCES

[1] 2023. Online Sexual Harassment. https://www.childnet.com/help-and-advice/online-sexual-harassment/.

[2] 2023. Sexual Harassment. https://www.eeoc.gov/sexual-harassment.

[3] Njood K Al-harbi and Amal A. Al-shargabi. 2023. An Exploratory Analysis of Using Chatbots in Academia. *International Journal of Advanced Computer Science and Applications* 14, 12 (2023). https://doi.org/10.14569/IJACSA.2023.0141212

[4] Ashwaq Alsoubai, Afsaneh Razi, Zainab Agha, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2024. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 114 (April 2024), 37 pages. https://doi.org/10.1145/3637391

[5] Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. From 'Friends with Benefits' to 'Sextortion:' A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 411 (Nov. 2022), 32 pages. https://doi.org/10.1145/3555136

[6] Ahmed E. Arafa, Rasha S. Elbahrawe, Nahed M. Saber, Safaa S. Ahmed, and Ahmed M. Abbas. 2017. Cyber Sexual Harassment: A Cross-Sectional Survey over Female University Students in Upper Egypt. *International Journal Of Community Medicine And Public Health* 5, 1 (Dec. 2017), 61. https://doi.org/10.18203/2394-6040.ijcmph20175763

[7] Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Chi '21)*. Association for Computing Machinery, New York, NY, USA, Article 257. https://doi.org/10.1145/3411764.3445318

[8] 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. (2022).   https://doi.org/10.48550/ARXIV.2212.08073

[9] A. Barak. SPR 2005. Sexual Harassment on the Internet. *SOCIAL SCIENCE COMPUTER REVIEW* 23, 1 (SPR 2005), 77–92.   https://doi.org/10.1177/0894439304271540

[10] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10).* Association for Computing Machinery, New York, NY, USA, 1301–1310.   https://doi.org/10.1145/1753326.1753521

[11] James Beckman. 2024. 120+ Chatbot Statistics for 2024 (Already Mainstream). *Techreport* (May 2024).

[12] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–19.   https://doi.org/10.1145/3134659

[13] Claire Boine. 2023. Emotional Attachment to AI Companions and European Law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2023 (Feb. 2023).   https://doi.org/10.21428/2c646de5.db67ec7f

[14] Virginia Braun and Victoria Clarke. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological.* American Psychological Association, Washington, DC, US, 57–71.   https://doi.org/10.1037/13620-004

[15] Elizabeth Broadbent. 2017. Interactions With Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology* 68, 1 (Jan. 2017), 627–652.   https://doi.org/10.1146/annurev-psych-010416-043958

[16] Amy Bruckman. 2002. Studying the Amateur Artist: A Perspective on Disguising Data Collected in Human Subjects Research on the Internet. *Ethics and Information Technology* 4 (2002), 217–231.

[17] Ankita Chakravarti. 2023. Replika AI Chatbot Stops Responding to Sexual Advances, Leaves Users Lonely and Lost. https://www.indiatoday.in/technology/news/story/replika-ai-chatbot-stops-responding-to-sexual-advances-leaves-users-lonely-and-lost-2333554-2023-02-16.

[18] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. 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 544, 20 pages.   https://doi.org/10.1145/3491102.3517475

[19] Raffaele Ciriello, Oliver Hannon, Angelina Chen, and Emmanuelle Vaast. 2023. Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika.

[20] Samantha Cole. 2023. 'My AI Is Sexually Harassing Me': Replika Users Say the Chatbot Has Gotten Way Too Horny. https://www.vice.com/en/article/z34d43/my-ai-is-sexually-harassing-me-replika-chatbot-nudes.

[21] Milica Cosic. 2023. My AI Is Sexually Harassing Me: Replika Users Say Chatbot Has Become Too Aroused. https://www.mirror.co.uk/news/world-news/ai-sexually-harassing-me-replika-29147565.

[22] Prema Dev, Jessica Medina, Zainab Agha, Munmun De Choudhury, Afsaneh Razi, and Pamela J. Wisniewski. 2022. From Ignoring Strangers' Solicitations to Mutual Sexting with Friends: Understanding Youth's Online Sexual Risks in Instagram Private Conversations. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, Taiwan) *(CSCW'22 Companion).* Association for Computing Machinery, New York, NY, USA, 94–97.   https://doi.org/10.1145/3500868.3559469

[23] Eleanor Drage, Kerry McInerney, and Jude Browne. 2024. Engineers on Responsibility: Feminist Approaches to Who's Responsible for Ethical AI. *Ethics and Information Technology* 26, 1 (March 2024), 4.   https://doi.org/10.1007/s10676-023-09739-1

[24] Maeve Duggan. 2017. *Online Harassment 2017.* Technical Report. Pew Research Center.

[25] Tessa Eagle, Conrad Blau, Sophie Bales, Noopur Desai, Victor Li, and Steve Whittaker. 2022. "I Don't Know What You Mean by 'I Am Anxious'": A New Method for Evaluating Conversational Agent Responses to Standardized Mental Health Inputs for Anxiety and Depression. *ACM Transactions on Interactive Intelligent Systems* 12, 2, Article 12 (July 2022).   https://doi.org/10.1145/3488057

[26] Denis Fedorenko. 2021. How We Moved from OpenAI.

[27] Louise F. Fitzgerald, Michele J. Gelfand, and Fritz Drasgow. 1995. Measuring Sexual Harassment: Theoretical and Psychometric Advances. *Basic and Applied Social Psychology* 17, 4 (Dec. 1995), 425–445.   https://doi.org/10.1207/s15324834basp1704_2

[28] Jaclyn Friedman and Jessica Valenti. 2019. *Yes Means Yes!: Visions of Female Sexual Power and a World without Rape.* Seal Press.

[29] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences* 120, 30 (July 2023), e2305016120.   https://doi.org/10.1073/pnas.

2305016120 arXiv:2303.15056 [cs]

[30] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving Alignment of Dialogue Agents via Targeted Human Judgements. https://doi.org/10.48550/arXiv.2209.14375 arXiv:2209.14375 [cs]

[31] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe Sexting: The Advice and Support Adolescents Receive from Peers Regarding Online Sexual Risks. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 42 (apr 2021), 31 pages. https://doi.org/10.1145/3449116

[32] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. 'If You Care About Me, You'll Send Me a Pic' - Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) *(CSCW '21).* Association for Computing Machinery, New York, NY, USA, 67–71. https://doi.org/10.1145/3462204.3481739

[33] Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* ACM, New Orleans LA USA, 123–129. https://doi.org/10.1145/3278721.3278777

[34] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. 2017. Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research* 27, 4 (March 2017), 591–608. https://doi.org/10.1177/1049732316665344

[35] Annabell Ho, Jeff Hancock, and Adam S Miner. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot. *Journal of Communication* 68, 4 (Aug. 2018), 712–733. https://doi.org/10.1093/joc/jqy026

[36] Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging ChatGPT for Efficient Fact-Checking. https://doi.org/10.31234/osf.io/qnjkf

[37] Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT Better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023.* 294–297. https://doi.org/10.1145/3543873.3587368 arXiv:2302.07736 [cs]

[38] Ziheng Huang, Sebastian Gutierrez, Hemanth Kamana, and Stephen Macneil. 2023. Memory Sandbox: Transparent and Interactive Memory Management for Conversational Agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* ACM, San Francisco CA USA, 1–3. https://doi.org/10.1145/3586182.3615796

[39] jackfromthesky. 2022. Replika App - Update Log.

[40] jackfromthesky. 2023. Update Log 2022.

[41] Vaishali Kaushal and Rajan Yadav. 2022. The Role of Chatbots in Academic Libraries: An Experience-based Perspective. *Journal of the Australian Library and Information Association* 71, 3 (July 2022), 215–232. https://doi.org/10.1080/24750158.2022.2106403

[42] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Chi '24).* Association for Computing Machinery, New York, NY, USA, Article 701. https://doi.org/10.1145/3613904.3642937

[43] Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification. arXiv:2303.03953 [cs]

[44] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too Human and Not Human Enough: A Grounded Theory Analysis of Mental Health Harms from Emotional Dependence on the Social Chatbot Replika. *New Media & Society* (Dec. 2022), 146144822111420. https://doi.org/10.1177/14614448221142007

[45] Molly K. Land and Jay D. Aronson. 2020. Human Rights and Technology: New Challenges for Justice and Accountability. *Annual Review of Law and Social Science* 16, 1 (Oct. 2020), 223–240. https://doi.org/10.1146/annurev-lawsocsci-060220-081955

[46] Qi Liu, Dani Yogatama, and Phil Blunsom. 2022. Relational Memory-Augmented Language Models. *Transactions of the Association for Computational Linguistics* 10 (May 2022), 555–572. https://doi.org/10.1162/tacl_a_00476

[47] Stefano De Luca. 2023. *New Product Liability Directive.* Briefing PE 739.341. European Parliamentary Research Service. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0495 EU Legislation in Progress.

[48] Jessica Lucas. 2024. The Teens Making Friends with AI Chatbots — Theverge.Com.

[49] Abigail Luger, Ewa; Sellen. 2016. CHI - "like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. , 5286–5297 pages. https://doi.org/10.1145/2858036.2858288

[50] Nick Lüthi, Christian Matt, Thomas Myrach, and Iris Junglas. 2023. Augmented Intelligence, Augmented Responsibility? *Business & Information Systems Engineering* 65, 4 (Aug. 2023), 391–401. https://doi.org/10.1007/s12599-023-00789-9

[51] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support.   https://doi.org/10.48550/arXiv.2307.15810 arXiv:2307.15810 [cs]

[52] Jingbo Meng, Minjin (MJ) Rheu, Yue Zhang, Yue Dai, and Wei Peng. 2023. Mediated Social Support for Distress Reduction: AI Chatbots vs. Human. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 72 (April 2023). https://doi.org/10.1145/3579505

[53] Cade Metz. 2020. Riding Out Quarantine With a Chatbot Friend: 'I Feel Very Connected'. *The New York Times* (June 2020).

[54] Muhammad Azam, Lubaina Zafar, Tanveer Rafiq, Sana Zafar, Umar Rafiq, and Muhammad Adnan. 2024. Enhancing Chatbot Intelligence Through Narrative Memory Structures. *The Asian Bulletin of Big Data Management* 4, 02 (May 2024). https://doi.org/10.62019/abbdm.v4i02.154

[55] Andreea Muresan and Henning Pohl. 2019. Chats with Bots: Balancing Imitation and Engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Chi Ea '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3313084

[56] Mohammad Namvarpour and Afsaneh Razi. 2024. Uncovering Contradictions in Human-AI Interactions: Lessons Learned from User Reviews of Replika. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, San Jose Costa Rica, 579–586. https://doi.org/10.1145/3678884.3681909

[57] Syeda Faiza Nasim, Muhammad Rizwan Ali, and Umme Kulsoom. 2022. Artificial Intelligence Incidents & Ethics A Narrative Review. *International Journal of Technology, Innovation and Management (IJTIM)* 2, 2 (Oct. 2022). https://doi.org/10.54489/ijtim.v2i2.80

[58] Gina Neff and Peter Nagy. 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10 (Oct. 2016), 4915–4931.

[59] Donald A. Norman. 2013. *The Design of Everyday Things* (revised and expanded edition ed.). Basic Books, New York, New York.

[60] Council of the European Communities. 1985. Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. , 29–33 pages. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31985L0374 Finnish special edition: Chapter 15 Volume 6 P. 0239; Spanish special edition: Chapter 13 Volume 19 P. 0008; Swedish special edition: Chapter 15 Volume 6 P. 0239; Portuguese special edition: Chapter 13 Volume 19 P. 0008.

[61] OpenAI. 2023. GPT-3.5-Turbo.

[62] OpenAI. 2023. GPT-4.

[63] Osiris1953. 2021. Update Regarding GPT.

[64] Namkee Park, Kyungeun Jang, Seonggyeol Cho, and Jinyoung Choi. 2021. Use of Offensive Language in Human-Artificial Intelligence Chatbot Interaction: The Effects of Ethical Ideology, Social Competence, and Perceived Human-likeness. *Computers in Human Behavior* 121 (Aug. 2021), 106795. https://doi.org/10.1016/j.chb.2021.106795

[65] European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various regulations and directives. http://data.europa.eu/eli/reg/2024/1689/oj Official Journal of the European Union, L 1689, 12.7.2024.

[66] 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 (March 2023), 107600. https://doi.org/10.1016/j.chb.2022.107600

[67] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa Is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[68] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 89 (April 2023), 29 pages. https://doi.org/10.1145/3579522

[69] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376400

[70] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (Oct. 2021), 38 pages. https://doi.org/10.1145/3479609

[71] Afsaneh Razi, John S. Seberger, Ashwaq Alsoubai, Nurun Naher, Munmun De Choudhury, and Pamela J. Wisniewski. 2024. Toward Trauma-Informed Research Practices with Youth in HCI: Caring for Participants and Research Assistants When Studying Sensitive Topics. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 134 (April 2024), 31 pages. https://doi.org/10.1145/3637411

[72] Michael V. Reiss. 2023. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. arXiv:2304.11085 [cs]

[73] Rizanna Rosemary, Arief Bayu Wardhana, Hamdani M. Syam, and Novi Susilawati. 2024. THE RELATIONSHIP BETWEEN ANONYMITY AND CYBER SEXUAL HARASSMENT BY TWITTER USERS: A CROSS-SECTIONAL STUDY. *Journal of Community Mental Health and Public Policy* 6, 2 (April 2024), 95–104. https://doi.org/10.51602/cmhp.v6i2.131

[74] Sigal Samuel. 2024. People Are Falling in Love with - and Getting Addicted to - AI Voices. *Vox* (Aug. 2024).

[75] Virginia Sapiro. 2018. Sexual Harassment: Performances of Gender, Sexuality, and Power. *Perspectives on Politics* 16, 4 (Dec. 2018), 1053–1066. https://doi.org/10.1017/S1537592718002815

[76] John S. Seberger, Irina Shklovski, Emily Swiatek, and Sameer Patil. 2022. Still Creepy After All These Years:The Normalization of Affective Discomfort in App Use. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3502112

[77] 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 (Dec. 2022), 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

[78] Phoebe Smith. 2023. If I Can't Predict My Future, Why Can AI? Exploring Human Interaction with Predictive Analytics. *Theses - ALL* (May 2023).

[79] Simon Ståhl and Inga Dennhag. 2021. Online and Offline Sexual Harassment Associations of Anxiety and Depression in an Adolescent Sample. *Nordic Journal of Psychiatry* 75, 5 (July 2021), 330–335. https://doi.org/10.1080/08039488.2020.1856924

[80] Stephanie. 2014. Cohen's Kappa Statistic. https://www.statisticshowto.com/cohens-kappa-statistic/.

[81] Yolande Strengers, Jathan Sadowski, Zhuying Li, Anna Shimshak, and Florian 'Floyd' Mueller. 2021. What Can HCI Learn from Sexual Consent? A Feminist Process of Embodied Consent for Interactions with Emerging Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445107

[82] Konstantin R. Strömel, Stanislas Henry, Tim Johansson, Jasmin Niess, and Paweł W. Woźniak. 2024. Narrating Fitness: Leveraging Large Language Models for Reflective Fitness Tracker Data Interpretation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Chi '24)*. Association for Computing Machinery, New York, NY, USA, Article 646. https://doi.org/10.1145/3613904.3642032

[83] Singgih Subiyantoro, I Nyoman Sudana Degeng, Dedi Kuswandi, and Saida Ulfa. 2023. Exploring the Impact of AI-Powered Chatbots (Chat GPT) on Education: A Qualitative Study on Benefits and Drawbacks. *Jurnal Pekommas* 8, 2 (Dec. 2023), 157–168. https://doi.org/10.56873/jpkm.v8i2.5206

[84] Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. User Expectations of Conversational Chatbots Based on Online Reviews. In *Designing Interactive Systems Conference 2021 (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1481–1491. https://doi.org/10.1145/3461778.3462125

[85] Vivian P. Ta-Johnson, Carolynn Boatfield, Xinyu Wang, Esther DeCero, Isabel C. Krupica, Sophie D. Rasof, Amelie Motzer, and Wiktoria M. Pedryc. 2022. Assessing the Topics and Motivating Factors Behind Human-Social Chatbot Interactions: Thematic Analysis of User Experiences. *JMIR Human Factors* 9, 4 (Oct. 2022), e38876. https://doi.org/10.2196/38876

[86] Caroline Tranberg. 2023. *"I Love My AI Girlfriend" A Study of Consent in AI-human Relationships*. Master's thesis. The University of Bergen.

[87] Tracy J. Trothen. 2022. Replika: Spiritual Enhancement Technology? *Religions* 13, 4 (April 2022), 275. https://doi.org/10.3390/rel13040275

[88] Mariek M P Vanden Abeele. 2021. Digital Wellbeing as a Dynamic Construct. *Communication Theory* 31, 4 (Nov. 2021), 932–955. https://doi.org/10.1093/ct/qtaa024

[89] Joana Varon and Paz Peña. 2021. Artificial Intelligence and Consent: A Feminist Anti-Colonial Critique. *Internet Policy Review* 10, 4 (Dec. 2021). https://doi.org/10.14763/2021.4.1602

[90] Venkatesh, Morris, Davis, and Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425. https://doi.org/10.2307/30036540 jstor:10.2307/30036540

[91] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1231–1245. https://doi.org/10.1145/2998181.2998337

[92] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and Social Risks of Harm from Language Models. (2021). https://doi.org/10.48550/ARXIV.2112.04359

[93] Tianling Xie and Iryna Pentina. 2022. *Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika.*

[94] Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The Role of AI in Peer Support for Young People: A Study of Preferences for Human- and AI-Generated Responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1006, 18 pages. https://doi.org/10.1145/3613904.3642574

[95] Bowen Zhang, Daijun Ding, and Liwen Jing. 2023. How Would Stance Detection Techniques Evolve after the Launch of ChatGPT? arXiv:2212.14548 [cs]

[96] Anne Zimmerman, Joel Janhonen, and Emily Beer. 2023. Human/AI Relationships: Challenges, Downsides, and Impacts on Human/Human Relationships. *AI and Ethics* (Oct. 2023). https://doi.org/10.1007/s43681-023-00348-8