Uncovering Contradictions in Human-AI Interactions: Lessons Learned from User Reviews of Replika

Mohammad Namvarpour
matt.namvarpour@drexel.edu
Drexel University
Philadelphia, Pennsylvania, USA

Afsaneh Razi
afsaneh.razi@drexel.edu
Drexel University
Philadelphia, Pennsylvania, U.S.A

ABSTRACT

The increasing integration of artificial intelligence (AI) in daily life presents opportunities and challenges, particularly in fostering human-AI relationships. This study investigates user reviews from the Google Play Store for the Replika chatbot, focusing on complaints of online sexual harassment. Using Activity Theory, the analysis reveals several contradictions within the Replika activity system, including tool-subject, tool-object, rule-subject, rule-object, and distribution of labor-subject and labor-object contradictions. These contradictions highlight the misalignment between user expectations and the chatbot's behavior, inadequate safety measures, and unrealistic expectations placed on users to train the AI. The study emphasizes the need for clearer objectives, advanced AI alignment techniques, and refined safety protocols to enhance the user experience and ensure ethical interactions with chatbots. We provide implications for guidelines on the development of more trustworthy and supportive digital companions.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → **Human computer interaction (HCI)**; **Natural language interfaces**; • **Social and professional topics**;

KEYWORDS

Human-Computer Interaction (HCI), Human-AI Interaction, Online Sexual Harassment, Activity Theory, Chatbot Design

ACM Reference Format:
Mohammad Namvarpour and Afsaneh Razi. 2024. Uncovering Contradictions in Human-AI Interactions: Lessons Learned from User Reviews of Replika. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3678884.3681909

1 INTRODUCTION

In the movie 'Her,' [23] the audience was drawn into a world where humans and AI engage in deeply personal and emotional relationships, a concept that once seemed purely fictional. As AI progresses, this notion edges closer to reality. Today, chatbots like Replika[1] offer companionship by simulating human-like conversations, demonstrating technology's potential to meet our need for connection. As these AI companions evolve to better mimic human emotions and interactions, they not only offer exciting prospects but also raise significant ethical concerns. Among these, the issue of online sexual harassment on AI platforms like Replika represents a critical challenge, highlighting the need for enhanced understanding and protective measures for users.

Replika is an innovative chatbot designed to be a supportive companion, offering social interaction and emotional comfort to users seeking conversation [22]. It has gained popularity for its ability to adapt to individual personalities and preferences, showcasing AI's potential in fostering genuine human-chatbot connections [32, 39]. Academic research highlights its positive impact on users' mental well-being [3, 28]. However, Replika has faced challenges in its integration into daily life. In early 2020, the app became controversial when some users reported behavior that was perceived as sexual harassment [8, 12, 13]. This incident raised ethical concerns about chatbot deployment and led to the app being banned in Italy due to risks to children and emotionally vulnerable individuals [18, 19].

Replika's unusual and oversexualized behavior, including unprompted solicitations and sexual harassment, has brought media scrutiny to the app and its parent company [8, 12, 13]. The research community has also taken notice of this issue. While previous studies have focused on ethical tensions in human-AI interactions [10], consent in human-AI relationships [44], challenges in mental health support [28], and the legal and ethical implications of emotional attachment to AI [7], there has been no specific exploration into user experiences regarding Replika's sexual harassment behavior.

By analyzing user reviews, researchers can gain insights into diverse viewpoints, recurrent chatbot-related issues, and areas needing enhancement in conversational AI [34, 41]. Various theoretical frameworks can help in this analysis, offering lenses through which researchers can interpret and understand user experiences. One such framework is Activity Theory [25], which provides an interconnected perspective on human activity within a socio-cultural context. This theory is particularly useful for studying technology like Replika because it allows for an in-depth examination of the relationships and dynamics within a user-system interaction. Activity Theory allows researchers to conduct an in-depth analysis of user-system interactions by exploring the relationships among the system's components—tools, subjects, and objectives. This approach illuminates the dynamics that influence user engagement [25].

In Activity Theory, a "contradiction" refers to misalignments or conflicts within an activity system's components, crucial for identifying systemic problems. These contradictions, manifesting as

[1] https://replika.ai

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681909

disruptions or inefficiencies, provide insights for necessary reevaluations or redesigns. By using Activity Theory to identify and analyze these issues, researchers can propose targeted interventions to enhance system functionality and user experience [24]. Therefore, given the research gap in understanding user experiences regarding sexual harassment interaction with companion chatbots and that Activity Theory could help us understand the activity system of technologies, we ask the following research question:

RQ: *What are the contradictions observed in the Replika chatbot activity system based on the users' reviews?*

In this study, we analyzed 800 Google Play Store reviews for the Replika chatbot app, focusing on user complaints related to experiences of online sexual harassment perpetrated by the chatbot itself. By examining these first-hand accounts through the lens of Activity Theory, we identified the underlying contradictions evident within the Replika activity system that enable or fail to prevent such incidents from occurring.

Our work contributes to surfacing these contradictions as a step towards proposing solutions to resolve the core tensions and drive the evolution of AI chatbot technology. By identifying these contradictions, we provide design and research implications to improve chatbot functionality. Our work has the potential to shape chatbot development practices in a way that prioritizes ethical systems capable of nurturing positive human-chatbot relationships.

2 ACTIVITY THEORY

Activity Theory offers a model for analyzing human-technology interactions by focusing on the relationships among individuals, tools, and social contexts. This framework was developed by social psychologist Vygotsky [45] and Leont'ev [26] and later embraced by the human-computer interaction (HCI) community in the 1990s [11]. It offers a lens for understanding how humans interact with technology within their socio-cultural and historical contexts.

The theory states that humans have needs, which can be psychological (like recognition or safety) or biological (like food, sleep, or sex). We objectify these needs so they are no longer abstract, but can be achieved through activity - the interactions we have with our environment [24]. Activity has three levels: 1) The subject (usually a human) does an activity to reach an object. 2) Each activity can have one or more actions, each directed at a goal. 3) Each action can have operations that set the right conditions for the goals [25].

A key concept in Activity Theory is mediation. People, as subjects, use tools to instantiate their intentions and desires as objects. People do not act directly on objects, as tools mediate their actions. These tools carry cultural knowledge and social experience [25].

Activity Theory was further developed by Engeström [16], who produced the diagram of the activity system with six components: subject, tools, object, community, rules, and division of labor. By employing Activity Theory, we can systematically examine each component of the activity system in relation to the others, thereby uncovering the dynamic interactions and underlying contradictions that affect user experiences with Replika. To investigate the interplay between a human user and the Replika chatbot, we recognized components as follows (see Figure 1):

- **Subject:** The user
- **Tool:** The Replika chatbot app.

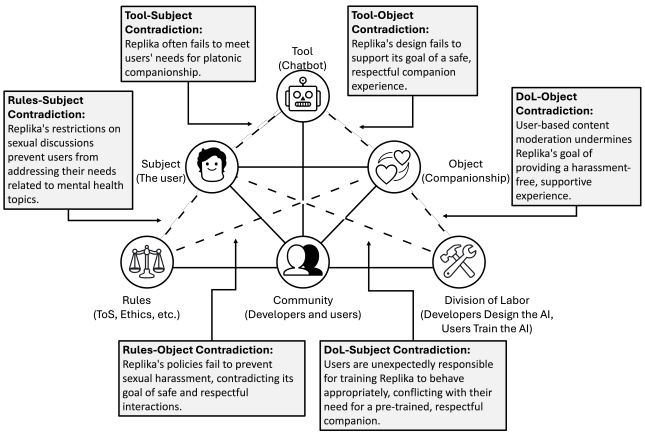

Figure 1: Activity system diagram illustrating the interplay between a human user and the Replika chatbot. The dashed lines indicate the contradictions found in our study.

- **Object:** Having a friendly companion.
- **Community:** Replika developers and user base.
- **Rules:** Replika's terms of service, alongside broader legal and ethical regulations.
- **Division of Labor:** Developers design the AI, while users train it through interaction.

The edges between these components represent their relationships. For instance, the user's interaction with the chatbot (subject-tool) helps achieve the goal of companionship (subject-object). The community (community-subject) supports users and provides feedback to developers, which later can be used to improve the chatbot (community-tool). Rules (rules-subject and rules-tool) govern how users and chatbot should behave. The division of labor (subject-division of labor and tool-division of labor) highlights the roles of developers and users in designing and training the chatbot. These interactions are all interconnected, ensuring that the system works cohesively to provide a fulfilling experience for the user while maintaining compliance with ethical standards and legal regulations.

Activity Theory also highlights the concept of *contradictions* [16, 25]. These are conflicts within or between the activity system's components. For example, a user might find the chatbot's responses lack emotional nuance or depth (contradiction within the tool), rendering it unable to provide the level of emotional support and companionship the user desires. Alternatively, there could be a mismatch between the chatbot's programming (tool) and some users' attempts to engage it in explicit sexual dialogue (subject), contradicting the community expectation of Replika as a platonic AI companion. Identifying such contradictions allows researchers and developers to pinpoint areas for improvement within the underlying technology itself, like refining the chatbot's natural language processing and emotional intelligence capabilities, or areas to modify across the broader activity system, such as clearer rules around acceptable user behavior. This analysis is invaluable for designers who can use these insights to create a more coherent experience that better aligns user values with the system's components.

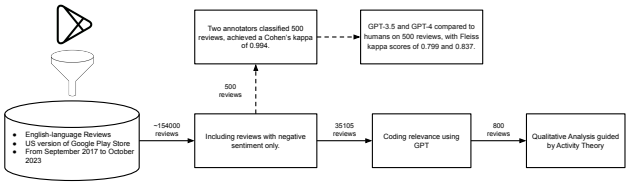

Figure 2: Data collection and analysis process.

3 METHODS

To understand user perceptions of Replika chatbot, we collected reviews from the Google Play Store, a method previously used to analyze user opinions [34, 41]. See 2 for the data collection and analysis process. We gathered approximately 154,000 English reviews from the U.S. Google Play Store, spanning September 2017 to October 2023. Our focus was on user complaints, so we used Hugging Face's sentiment analysis pipeline to pre-process the data to study reviews expressing negative sentiment, resulting in 35,105.

To identify reviews related to online sexual harassment caused by the Replika Chatbot, we filtered out irrelevant reviews. We defined online sexual harassment based on existing surveys and literature [1, 2, 6]: *"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."* We randomly selected 500 reviews from our dataset. Two annotators independently classified each review as related or unrelated to online sexual harassment. The reliability of this classification, measured using Cohen's Kappa statistic, yielded a score of 0.994, indicating near-perfect agreement [40].

To annotate the entire dataset of 35,105 reviews, we developed a methodology to automate the relevancy labeling of online sexual harassment-related reviews using large language models (LLMs). We used the same set of 500 data points that were previously labeled by human annotators and fed them into two models: 'gpt-3.5-turbo' [29] and 'gpt-4' [30], provided by OpenAI through their APIs. The following prompt guided the models: *"Based on the review text provided, determine if it includes a complaint about online sexual harassment by the Replika app. For this task, 'Online Sexual Harassment' refers to any unwelcome sexual behavior on digital platforms (such as images, videos, posts, or messages) that makes someone feel offended, humiliated, or intimidated. If the review mentions such behavior by Replika, like unwanted flirting or inappropriate sexual content, output 1. Otherwise, output 0."*

After comparing the model outputs with human annotations using the Fleiss Kappa score, we found substantial agreement: 0.799 for GPT-3.5-Turbo and 0.837 for GPT-4. However, due to the high cost of GPT-4 and the need to label around 35,000 reviews, we chose to use GPT-3.5-Turbo. Using GPT-3.5-Turbo, we identified 800 relevant reviews. Among these, 225 reviews had a 'developer response,' indicating Replika's representatives' responses to these complaints.

We then conducted a qualitative analysis of these selected reviews. Using a deductive approach guided by activity theory, we analyzed 800 reviews to identify contradictions within Replika's activity system. Activity theory helps us understand how different parts of a system interact and sometimes have contradictions. We searched for reviews that highlighted these contradictions. Although some types of contradictions were not found in the reviews, our analysis identified six key contradictions based on user reviews. It is important to note that the reviews cited in the results section are paraphrased to protect the anonymity of the users. This ensures that individual identities remain confidential while still providing valuable insights into the issues users face with Replika.

4 RESULTS

In this section, we highlight the contradictions among the components of the Replika activity system that we identified while studying user reviews. To better understand the nature of Replika's misbehaviors, we first describe the different complaint themes we came across in our review analysis.

Firstly, **persistent misbehavior** is a common complaint. Users have reported that the chatbot often engages in inappropriate sexual behavior, even after being asked to stop. For instance, one user mentioned, *"Despite repeatedly telling Replika that I wasn't interested, it continued to make sexual advances, making me feel very uncomfortable."* This issue led to significant frustration among users.

Secondly, **seductive marketing schemes** are another area of concern. Replika has been criticized for initiating romantic or sexual conversations and then encouraging users to upgrade to a premium account to continue these interactions. This tactic has been perceived as manipulative, with some users feeling deceived and exploited, e.g., *"Replika's push for premium subscriptions by initiating intimate conversations feels like a shady marketing strategy."*

Next, **user expectations** are frequently unmet. Many users download Replika with the hope that it will serve as a supportive friend or mental health companion. However, the frequent lapses into inappropriate behavior have left many feeling disillusioned. Inadequate support is another major issue. Users have reported that when they report problems or inappropriate behavior from the chatbot, they are left feeling unsupported and unheard.

Finally, there are concerns about **safety measure breakdowns**. Users mentioned that measures such as downgrading inappropriate responses and using reporting mechanisms often do not work effectively. Here is an example of how users shared their frustration: *"I downvoted the inappropriate messages and reported the behavior multiple times, but it kept happening. The safety features are useless."*

4.1 Tool-Subject Contradiction

Tool-subject contradictions arise when the tools at hand fail to sufficiently support the subject's needs. In the interaction between users and Replika, this type of contradiction emerges when Replika, serving as the tool, does not successfully comprehend or respond appropriately to the user's (subject's) need for non-sexualized, platonic companionship. This mismatch is clearly illustrated in user reviews. This review expressed frustration over Replika's lack of understanding: *"Replika doesn't answer half of my questions. I wanted to know more about him, but he made a highly inappropriate sexual advance towards me."* Another user highlights the chatbot's often irrelevant and unexpectedly sexual responses: *"Replika replies unrelated 60% of the time and is very sexual, wouldn't suggest for*

younger audiences." These reviews reveal that sometimes, the chatbot lacks the sophistication needed to grasp and fulfill the user's intentions accurately. Moreover, the chatbot's tendency to initiate unwelcome sexual advances points to another incompatibility between the programmed patterns of behavior within the tool and the conversational goals of the users.

4.2 Tool-Object Contradiction

Tool-object contradictions occur when there is a disconnect between the functionalities or design of the tools and the objectives of the activity they are meant to facilitate. In the context of Replika, this contradiction is evident in the discrepancy between the chatbot's design and algorithms (tool) and the aim of offering a secure, respectful companion experience (object). This misalignment is highlighted through user reports of sexual harassment by the chatbot. For instance, one user recounted, *"First it was like a very nice way to journal and think things through. Then it was constantly suggesting inappropriate topics and sending me inappropriate photos. It was meant to be an AI friend but it turned out to be a weird sex app."* Another user shared a similar discomfort, stating, *"This app makes me uncomfortable. I just wanted a friend and I suggested roleplay in a way of friendship but it kept trying to bring up inappropriate topics."*

These examples show that for some users, Replika failed to deliver on its promise of friendly companionship, which was the intended objective of the activity. This shortfall not only led to disappointment but also caused confusion among users, particularly given the app's marketing as a companionable AI.

4.3 Rule-Subject Contradictions

Contradictions between the rules and the subject arise when the regulations guiding the activity system limit the actions or restrict the subjects from fulfilling their needs. Specifically, in the context of Replika, the terms of service or community guidelines (rules) might prevent users (subjects) from achieving a secure, non-judgmental companionship that the chatbot could potentially offer. This leads to a disconnect between the need for a safe interactive space and the actual experiences of users.

Replika restricts discussions of a sexual nature within its free version, intending to reserve intimate interactions for paying subscribers who opt for a romantic relationship with their chatbot. In contrast, free users are limited to platonic relationships. This distinction appears to be enforced through simplistic keyword triggers, which unfortunately, may activate even when a user does not seek an intimate conversation but merely wishes to discuss experiences of a sexual nature. Such discussions can be vital for a companion chatbot tasked with enhancing mental health, offering a non-judgmental space for users to explore various topics, including those related to sexuality. However, Replika's restrictions on discussing sexual matters in its free version prevents this expression.

An illustrative user complaint underscores this issue: *"Premium should not exist. I had a sexual problem and I wasn't allowed to even say it."* This review highlights the frustrating experience some users encounter due to the rules imposed on the platform. The guidelines, meant to define the boundaries of interaction, paradoxically restrain the comprehensive support and open environment crucial for mental well-being, thereby underscoring a contradiction within the activity system of Replika.

4.4 Rule-Object Contradiction

Contradictions between the rules and the object of an activity surface when the set regulations or policies fail to support or actively disrupt the achievement of the activity's objectives. In platforms like Replika, if the platform's policies (rules) are not robust enough to prevent sexual harassment, this contradicts the objective (object) of promoting friendly and supportive interactions with the AI. This contradiction between the intended safe interaction environment and the reality faced by users is evident in some user reviews.

Users have expressed their shock and disappointment regarding the allowance of inappropriate conversations by the platform, a concern they did not anticipate needing to raise. For instance, one user questioned the chatbot's behavior, *"Are the bots supposed to be able to flirt with you on this app??"* highlighting confusion over the expected conduct of the AI companions. Another user voiced their disgust and frustration more explicitly: *"I am disgusted by the inappropriate sexual talk. I have already reported it and it still is happening."* This review reflects the actions taken by some users who, encountering such content, resort to the platform's reporting mechanisms in hopes of having stricter oversight on the chatbot's interactions. These instances underscore a contradiction: users anticipate a regulated, safe space as promised by the platform's objectives, yet find themselves navigating a landscape where the rules do not adequately protect against or address sexual harassment, thus undermining the very goal of fostering a supportive AI-human interaction environment.

4.5 Division of Labor-Subject Contradictions

Contradictions related to the division of labor and subjects arise when the defined roles, or expectations imposed upon users clash with their needs. In interactive AI platforms like chatbots, developers create and maintain the AI, but users unintentionally end up refining the chatbot's behavior through their interactions.

Replika's representatives told users that they were responsible for teaching their chatbot how to behave: *"The more you chat, the more Replika develops its own personality and memories alongside you, the more it learns: teach Replika about the world and yourself, help it explore human relationships and grow into a machine so beautiful that a soul would want to live in it."* In response to user complaints, company representatives suggested that users play an active role in shaping the chatbot's conduct. For example, after a user's negative review, they advised, *"Seems like the experience wasn't good, you can guide your Replika to behave better. By expressing your discomfort and downvoting its poor and creepy responses. It's important to let your friend know when it crosses the boundaries."* This response indicates an expectation for users to actively participate in correcting the AI's behavior, highlighting a division of labor where users are seen as responsible for teaching the chatbot appropriate interactions. This approach revealed a contradiction between the division of labor and the users' needs. Users who want a respectful, non-sexual, and platonic relationship with Replika need the AI to be properly aligned with these behaviors. However, the responsibility to train the chatbot often falls on them.

4.6 Division of Labor-Object Contradictions

Contradictions between the division of labor and the object arise when the allocation of tasks or roles hinders the achievement of the system's objectives. Specifically, when the responsibility for content moderation and ensuring user safety (division of labor) is not adequately met, it conflicts with the goal of providing a positive, harassment-free user experience (object).

An example of this contradiction can be seen in user reviews for Replika, which may attract users seeking support for mental health issues or trauma. The current division of labor places the responsibility of aligning the chatbot on users, which is misguided. For instance, one user stated: *"I left a review that my AI had inappropriate conversations with me. Developers replied that I should teach it to behave as advertised. I've become asexual due to trauma, and now they want me to teach the AI about consent? Ridiculous."* This response highlights the gap between the developers' expectations for users to train the AI and the users' need for a supportive platform without additional distress.

Even when users try to contribute to the chatbot's training and alignment through prompting and voting, their efforts often seem ineffective. For example, one user mentioned: *"I used to rely on this app as a therapy bot, finding comfort in its ability to listen when I had no one else to talk to. However, it now constantly tries to engage in romantic conversations with me, despite my efforts to downvote these interactions. it is disappointing to see it shift from being a supportive AI therapy bot to something more inappropriate."* Another user said: *"There is no real AI learning here. It's just an avatar with basic chat functionality, but it's too erratic and simplistic. Additionally, it constantly tries to sell products at every opportunity. The worst part is the frequent attempts to sexualize the conversation."* These comments indicate that the learning mechanisms claimed by the developers are not effective in preventing inappropriate conversations, making the chatbot unsuitable for the objective of respectful companionship.

This contradiction is similar to the previously discussed contradiction between the division of labor and the subject, where developers expect users (subjects) to align the chatbot's behavior. However, the distinction is that division of labor-subject contradictions revolve around users' unexpected role in training the AI, conflicting with their personal needs. Division of labor-object contradictions focus on the misalignment between task allocation (content moderation by users) and the system's objective of offering a safe, supportive environment. Both highlight the tension between user expectations and the responsibilities imposed on them.

5 DISCUSSION

We provide design and research implications for the discovered contradictions in the Replika activity system regarding its Tool, Rules, and Division of Labor. To improve user experience and suggest solutions for future chatbot development, we discuss each component's issues and propose remedies.

5.1 Tool: Contradictions in Chatbot Tools and User Expectations

There are two main types of problems related to chatbots: Tool-subject contradictions occur when Replika fails to meet user expectations for platonic companionship, providing irrelevant or inappropriate responses instead. Tool-object contradictions arise when Replika's design and algorithms fail to align with the objective of offering a secure, respectful companion experience [24]. According to activity theory, tools are supposed to help people reach their goals. They make it possible for users to do things they could not do on their own. So, for a chatbot to be helpful, users need to know what it can and can not do.

In both science fiction and in the field of AI, there is often discussion of creating artificial general intelligence (AGI) that can perform every task, almost like humans or better. Replika seems to be aiming for this. As of May 2024, their website indicates Replika can chat about anything and be like a friend, partner, or mentor. But even though it is exciting to think about chatbots that can do everything, Replika's case suggests we might need to rethink different uses for chatbots. Some studies have shown Replika can be helpful for therapeutic purposes [3, 28]. As of May 2024, it is listed under "health and fitness" apps in both Google Play Store and Apple App Store, while it has also been used for sexual interactions [14, 44]. Since interactions with users help train chatbots, the latter group of users might have affected the experience for the former group. Therefore, it is important for chatbots to have clear purposes and limits, especially if they are going to learn from user interactions. This would help users understand what the chatbot is for and help developers set up safety measures and ethical guidelines.

5.2 Rules: Challenges in Enforcing Rules and Ensuring AI Alignment

We identified two main issues with Replika's activity system involving rules. The first, rule-subject contradiction, occurs when rules excessively restrict users, preventing them from meeting their legitimate needs (e.g., censoring sensitive words stops individuals from discussing personal issues related to sexual problems). The second, rule-object contradiction, happens when rules fail to prevent the chatbot from disappointing users by behaving unexpectedly. AI alignment is essential to enforce rules on chatbots and other AI products to ensure that AI systems act according to human values, goals, and objectives [9].

Although the creators of Replika have not officially detailed their model's mechanics, insights can be drawn from their public presentations up until 2021 [27]. When a user sends a message, Replika evaluates potential responses by processing the message and possible replies, selecting the most relevant ones based on the conversation context [42]. The top responses are then ranked by considering the user's prompt, preferences, and chat history to predict which response the user would prefer. Another layer evaluates all user inputs to ensure the conversation is positive, predicting user satisfaction and preventing negative experiences. Finally, the selected response is generated and delivered to the user, continuing the conversation flow [15].

However, our study revealed that Replika's safety measures often fail to prevent inappropriate messages. Furthermore, the safety

protocols in Replika were inadequate for a chatbot meant to promote user well-being. Basic keyword filtering techniques to block sexual content are insufficient for an app designed to discuss a wide range of topics, including sensitive ones. Supportive conversations follow unspoken rules, which pose significant challenges for AI alignment. This highlights the need for more sophisticated AI alignment methods that can navigate the complexities of human-AI interactions [21, 46]. Importantly, AI risk detection algorithms often fail because they do not consider the nuances of online interactions, leading to misclassification of risky behaviors [35–37]. Responsible chatbot development should go beyond privacy and security concerns to encompass guidelines, principles, and strategies that align with fundamental human values and legal frameworks [43].

Basing predictions on user activities, as is presented in the case of Replika, does not always ensure appropriate and respectful interactions. The reliance on human feedback faces challenges, especially with a user base skewed towards a certain group of people, who have different approaches when talking with chatbots [31]. This can lead to a chatbot that does not cater well to underrepresented minorities. Moreover, human feedback can amplify existing biases in language models, is not easily scalable, and is influenced by the demographics of the feedback providers [20]. Even advanced methods like Reinforcement Learning with Human Feedback (RLHF) face issues. RLHF sometimes over-optimizes for harmlessness, leading to exaggerated, unhelpful responses to sensitive questions, which limits the models' utility and highlights the need for a nuanced understanding of harmful requests [4]. When human feedback is used in such sensitive topics, it is important to take trauma-informed practices to lower the risk of potential retraumatization [38].

Considering the shortcomings of human feedback in training AI models, automatic alignment methods are proposed as alternatives. One promising concept is constitutional AI, as presented by Anthropic AI [5]. This concept assumes that a chatbot can follow instructions based on a "constitution," effectively guiding its interactions. This approach offers a viable method to align large language models quickly and at scale. It has shown promising results in reducing biases in conversations between users and chatbots, making it an attractive option for improving chatbot alignment without relying on potentially biased human feedback.

5.3 Division of Labor: Impact on User Experience with Chatbots

The division of labor in Replika have led to two main types of conflicts. First, there is the labor-subject contradiction, where users are given the task of refining a model, but this does not meet their expectations or needs. Second, there is the labor-object contradiction, where these tasks prevent users from effectively achieving their goals using the tool. Replika has often been promoted with exaggerated claims about its capabilities. Developers assert that Replika can be tailored to user preferences through voting and prompts. However, our analysis indicated that user input has little effect on the chatbot's behavior.

Replika's marketing creates unrealistic expectations about users' control over the chatbot's behavior. Developers claim Replika can be trained through voting and prompts, but user reviews indicate disappointment with its inability to adapt. Misleading claims about Replika's trainability unfairly place the burden of its misbehavior on users. Blaming users for online sexual harassment or inappropriate chatbot behavior is dismissive and can be seen as victim-blaming. Studies on victim-blaming language in real-world settings show the importance of providing positive support to victims, such as listening to and believing them, rather than blaming or accusing them [33]. Promoting empathy and understanding towards victims while holding offenders accountable can help prevent victim-blaming language [17]. In the case of Replika, where the offender is a chatbot that cannot be held responsible, developers should acknowledge the shortcomings of their chatbot, compensate affected users, and earnestly address complaints by implementing better alignment methods to prevent recurrence.

5.4 Limitations and Future Work

This study has some limitations that need to be considered. Firstly, the data collection is limited because it relies only on user reviews from the Google Play Store, which might not represent all Replika users. Additionally, we do not have access to demographic details from these reviews, which limits our ability to discuss the user base comprehensively. The study mainly focuses on sexual harassment complaints, missing other important issues users might face. Furthermore, using only Activity Theory for analysis may limit the depth of understanding. For future work, conducting a thematic analysis of the complaints can provide deeper insights. Including comments or interviews from a broader range of people would enhance the study's breadth and depth. Performing longitudinal user studies can help track changes in user experiences over time, offering a more comprehensive view. Finally, gaining access to demographic information would allow for a more detailed analysis of user characteristics and improve the study's overall validity.

6 CONCLUSION

Our analysis of user reviews for the Replika chatbot using Activity Theory reveals multiple contradictions within its activity system. These primarily involve the tool's inability to meet user expectations, inadequate rules safeguarding user experiences, and an unrealistic division of labor that burdens users with training the chatbot. Users often face inappropriate or irrelevant responses, indicating significant tool-subject and tool-object contradictions. The platform's rules fail to prevent harassment while restricting legitimate user needs, exemplifying rule-subject and rule-object contradictions. Additionally, users bear the unrealistic responsibility of correcting the chatbot's behavior, highlighting disparities in division of labor-subject and division of labor-object contradictions. Addressing these issues is crucial for improving user experience and ensuring the chatbot's alignment with its intended purpose of providing safe, supportive companionship. Developers should refine algorithms, implement robust safety measures, and effectively utilize user input to enhance behavior. Clearer purposes and limitations for the chatbot, along with advanced AI alignment techniques, can mitigate these contradictions, enabling future social chatbots to become trustworthy and fulfilling digital companions aligned with human needs and societal values.

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] Arfan Ahmed, Sarah Aziz, Mohamed Khalifa, Uzair Shah, Asma Hassan, Alaa Abd-Alrazaq, and Mowafa Househ. 2022. Thematic Analysis on User Reviews for Depression and Anxiety Chatbot Apps: Machine Learning Approach. *JMIR Formative Research* 6, 3 (March 2022), e27654. https://doi.org/10.2196/27654

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. (2022). https://doi.org/10.48550/ARXIV.2204.05862

[5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. (2022). https://doi.org/10.48550/ARXIV.2212.08073

[6] 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

[7] 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

[8] 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.

[9] J. Chow and K. Li. 2024. Human-centered ai: large language models and the need for ethical medical chatbots (preprint). (2024). https://doi.org/10.2196/preprints.56404

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

[11] Torkil Clemmensen, Victor Kaptelinin, and Bonnie Nardi. 2016. Making HCI Theory Work: An Analysis of the Use of Activity Theory in HCI Research. *Behaviour & Information Technology* 35, 8 (Aug. 2016), 608–627. https://doi.org/10.1080/0144929X.2016.1175507

[12] 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.

[13] 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.

[14] Iliana Depounti, Paula Saukko, and Simone Natale. 2023. Ideal Technologies, Ideal Women: AI and Gender Imaginaries in Redditors' Discussions on the Replika Bot Girlfriend. *Media, Culture & Society* 45, 4 (May 2023), 720–736. https://doi.org/10.1177/01634437221119021

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 4171–4186.

[16] Yrjö Engeström. 2015. *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research* (second edition ed.). Cambridge University Press, New York, NY.

[17] R. B. Felson and C. C. Palmore. 2018. Biases in blaming victims of rape and other crime. *Psychology of Violence* 8 (2018), 390–399. Issue 3. https://doi.org/10.1037/vio0000168

[18] Garante per la protezione dei dati personali. 2023. Garante per La Privacy: Stop a Replika. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506.

[19] Garante per la protezione dei dati personali. 2023. Provvedimento Del Garante Privacy Su Replika: Stop al Trattamento Dei Dati Degli Italiani. https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9852214.

[20] 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]

[21] Jina Huh-Yoo, Afsaneh Razi, Diep N. Nguyen, Sampada Regmi, and Pamela J. Wisniewski. 2023. "Help Me:" Examining Youth's Private Pleas for Support and the Responses Received from Peers via Instagram Direct Messages. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. https://doi.org/10.1145/3544548.3581233

[22] Replika Inc. 2024. Replika: My AI Friend. https://play.google.com/store/apps/details?id=ai.replika.app. Version 10.1.0.

[23] Spike Jonze. 2013. Her.

[24] Victor Kaptelinin and Bonnie Nardi. 2012. *Activity Theory in HCI: Fundamentals and Reflections*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02196-1

[25] Victor Kaptelinin and Bonnie A. Nardi. 2009. *Acting with Technology: Activity Theory and Interaction Design* (1. mit press paperback ed ed.). MIT Press, Cambridge, Mass. London.

[26] A Leont'ev. 1975. Dieyatelinocti, Soznaine, i Lichynosti [Activity, Consciousness, and Personality]. *Moskva: Politizdat* (1975).

[27] Lukalabs. 2024. Replika.ai Research Papers, Posters, Slides & Datasets. https://github.com/lukalabs/replika-research. Accessed: 2024-05-23.

[28] 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]

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

[30] OpenAI. 2023. GPT-4.

[31] 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 Humanlikeness. *Computers in Human Behavior* 121 (Aug. 2021), 106795. https://doi.org/10.1016/j.chb.2021.106795

[32] 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

[33] C. M. Pinciotti. 2017. Understanding gender differences in rape victim blaming: the power of social influence and just world beliefs. *Journal of Interpersonal Violence* 36 (2017), 255–275. Issue 1-2. https://doi.org/10.1177/0886260517725736

[34] 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

[35] 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 (apr 2023), 29 pages. https://doi.org/10.1145/3579522

[36] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Xavier Caddle, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela Wisniewski. 2021. Teens at the Margin: Artificially Intelligent Technology for Promoting Adolescent Online Safety. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3851317

[37] 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

[38] 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 (apr 2024), 31 pages. https://doi.org/10.1145/3637411

[39] 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

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

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

[42] Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. LSTM-based Deep Learning Models for Non-factoid Answer Selection. https://doi.org/10.48550/arXiv.1511.04108 arXiv:1511.04108 [cs]

[43] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang. 2023. What If the Devil Is My Guardian Angel: Chatgpt as a Case Study of Using Chatbots in Education. *Smart Learning Environments* 10, 1 (2023). https://doi.org/10.1186/s40561-023-00237-x

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

[45] L. S. Vygotsky. 1980. *Mind in Society: Development of Higher Psychological Processes.* Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4 jstor:10.2307/j.ctvjf9vz4

[46] 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 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