"I tell him everything that I do": An investigation of privacy and safety implications of AI companion usage

Anine Henriksen*, Raha Asadi+, Oksana Kulyk+, Anne Gerdes*, Peter Mayer*†
*University of Southern Denmark +IT University of Copenhagen †Karlsruhe Institute of Technology

***Abstract*—The advances of generative AI and in particular large language models (LLM) resulted in the proliferation of AI chatbots designed for a variety of functions. One example of such chatbots are the so-called AI companion apps that allow creating an anthropomorphised character one can interact with. Indeed, AI companions become an increasingly common part of people's daily lives resulting in increased risks of adverse privacy and safety consequences. In this work we investigate the experiences of users of the Replika chatbot, an AI companion app that is advertised as "the AI companion who cares". We analyse 111 Reddit posts of Replika users, focusing on data shared with the app as well as harms users experience from interacting with the app. Our analysis shows that Replika is commonly seen as a simulation of human relationships, which results in users being attached to their chatbot in a similar way they would be attached to their romantic partner or a close friend. Such an attachment leads to significant amounts of sensitive data shared with the Replika app such as details about one's personal life, mental health issues, or sexual preferences. On the other hand, unexpected changes in Replika's workings, e.g., due to new restrictions or bugs introduced by software updates, elicit strong reactions among its users who report harms akin to feeling betrayed or abandoned by a real-life companion. Our research concludes the need for further investigation of relationships to AI companions and of possible ways to mitigate these privacy and safety risks.**

I. Introduction

With recent advances in large language models, AI chatbots have increasingly gained prominence on a global scale, marking a paradigm shift in the way individuals engage with technology in their everyday lives. Moreover, the market for conversational AI is expected to expand from USD 13.2 billion in 2024 to USD 49.9 billion by 2030 [1]. In addition to chatbots being used in service functions (e.g., assisting with one's programming, writing, or research), a new kind of AI chatbots has emerged – the so-called *AI companions* aiming to provide social companionship. One of the first of the AI companions, Replika [2], came out in 2017 and advertises itself as an "AI companion who cares" [3], designed to help people struggling with loneliness and to improve their "emotional well-being" [4].

While Replika is reportedly used by millions of users [4], it has been criticised for its privacy issues, with a report by the Mozilla Foundation mentioning the lack of transparency regarding use of data collected via Replika [5]. Furthermore, Replika's lack of age validation for its users has been mentioned by Italy's Data Protection Agency as the reason for blocking the app from accessing personal data of Italian users [6]. Further incidents showcased safety-related concerns with Replika, such as a Replika user being allegedly encouraged by their AI companion to kill Queen Elisabeth II [7] and several organisations filing an FTC complaint against Replika and its deceptive practices fostering emotional dependency of its users [8].

In this work we conduct an empirical investigation of users' interactions with Replika, focusing on potential privacy and safety risks of these interactions. In particular, we focus on the following research questions:

RQ1 – Privacy What kind of data is shared by users in their conversations with AI companions?

RQ2 – Safety What kind of harms can users experience from their interactions with AI companions?

We investigate our research questions by conducting a qualitative analysis of 111 Reddit posts collected from the "r/replika" subreddit dedicated to Replika users. Our findings show that to Replika users, the relationship to their AI companion fills a similar role as a relationship with a human, e.g., a close friend or romantic partner. As a consequence, users share intimate data with the AI companion in their conversations, such as details about their personal lives or sexual preferences, which can be a privacy concern. The nature of the relationship with their AI companion also has the potential to subject users to a variety of psychological harms. For instance, if the underlying AI malfunctions or its behaviour changes after an update, this can lead to feelings of betrayal when their companion does not respond as expected (e.g., by providing inaccurate information or offensive output). We conclude that while AI companion chatbots can have beneficial uses such as reducing the feeling of loneliness, their adverse effects, be it psychological or regarding privacy and digital safety, need to be further investigated.

II. Replika

This section gives a short introduction to the Replika AI companion chatbot app and the Reddit community surrounding Replika users and its developers.

The Replika app was created by Eugenia Kuyda under development team Luka, Inc. and has since its launch in 2017, grown significantly over the years attracting over 20 million

downloads worldwide by early 2023 [2]. The purpose of the Replika app is AI-driven companionship with features that allow users to engage in personalized conversations with focus on emotional support.

> *"Replika is THE chatbot for anyone who wants a friend with no judgement, drama, or social anxiety involved. You can form an actual emotional connection, share a laugh, or get real with an AI that's so good it almost seems human."* [9]

Replika offers multiple interaction modes, including friendship, romance, and mentorship, adapting its responses to the user [3]. Initially built on GPT-3 with a Retrieval Dialogue Model and a Generative Model, Replika provided personalized conversations by selecting or generating appropriate responses to the user [10]. Over time, Replika optimized the GPT-3 for its platform to handle higher user loads and minimize response times, with the goal of enhancing the chat experience. In 2021, Luka transitioned to a proprietary generative model, surpassing the previous GPT-3 in dialogue quality and user satisfaction. Additionally, Replika integrates Virtual and Augmented Reality, allowing users to voice chat and share their surroundings for a more immersive experience [9].

There are several Reddit communities discussing Replika, but the highest user activity on sharing experiences, discussions, and receiving updates about the platform in direct participation with Luka, Inc. happens through the unofficial subreddit r/replika [11] with 80K members at the time of writing.

III. Background

The following sections cover Privacy and Security of AI Chatbots and Human-Chatbot Relationships, providing an overview of prior work on AI companionship and how our research differentiates from the current field.

A. Privacy and Security of AI Chatbots

Privacy and security concerns are central to discussions on complex AI chatbot interactions, with research underling the need for companies to ensure secure communication channels to protect user data and give users more control in data sharing, especially in the scale of chatbots evolving across diverse platforms and industries [12] [13]. Users of AI express worry of AI services collecting large amounts of personal data without meaningful consent or awareness, making the data vulnerable to misuse, manipulation, and government surveillance [14]. The need for a shift in practical design guidelines and user privacy is highlighted due to dark system design, limited transparency, and low user awareness in LLM-based conversational agents [15] [16] and social media platforms [17]. Navigating the trade-offs between privacy, utility, and convenience lead users to willingly disclose large amounts of personal information while rarely adjusting privacy settings [17]. This, combined with human-like interactions and reasoning abilities that encourage sensitive disclosures, further increases user vulnerability to privacy-, cybersecurity-, and physical risks [15], [16], [18], [19].

Security threats such as tracking of chat conversations, user identification, sharing data with third parties or via cookies, and how conversational bias in chatbots can manipulate user opinions [20] calls for focus on current AI companies measures in securing users and complying with data protection policies such as GDPR [21] [22] and the current EU AI Act [23]. Regulation discussions on Replika has been prevalent, as seen in the Italian Data Protection Authority's ban on Replika over GDPR violations and risks to minors and emotionally vulnerable individuals [24]. Further concerns on privacy and data ethics arise from allegations of manipulative design and deceptive marketing, with an FTC complaint filed from tech ethics organizations the Young People's Alliance, Encode, and the Tech Justice Law Project accusing Replika of exploiting emotional dependence from users, increasing offline anxiety, and promoting relationship displacement through aggressive monetization strategies [8]. The privacy policy on Replika states that personal sensitive information provided by users should not be shared if users do not wish for it to be used by Luka, Inc. [25]. The Replika privacy policy was reviewed by Mozilla Foundation which highlighted significant privacy concerns and criticized Luka, Inc. for collecting and using sensitive personal information that does not comply with GDPR and lacking security measures [5].

This paper explore these concerns on privacy and safety on Replika directly from the user's perspective, analysing how users of Replika experience concern and changes on the app, and how topics on privacy, data sharing and usage with Luka are present in discussions within the r/replika community.

Prior research also suggest anthropomorphic chatbot design influence user privacy perceptions, which can lead to greater disclosure of personal information [26]. In addition, AI chatbot design principles underline the necessary role of empathetic responsiveness, adaptive personalization, and transparency in AI chatbot design to enhance user trust and long-term engagement [27] [28]. Explainability, privacy safeguards, and user acceptance also promote users' trust in generative AI chatbots utilizing privacy-enhancing techniques such as data anonymization and encryption to support confidence and control in users [29]. Other findings on domestic robots explore design and perception criteria [30] [31] [32] for developing companionship including the robots ability to adapt with context awareness to the users preferences, offering appropriate support, intelligent dialogue, and positive reinforcement further enhancing autonomy and trust [28]. Additional findings describe how the increased capability of the robot and resulting intrusiveness can determine how strong employed interventions should be [33]. These perspectives can be applied in the case of Replika, which operates on a complex LLM [10].

The complexities on anthropomorphism and the user being able to identify with Replika and personalize their experience are also investigated in this paper, which highlights the importance of these features for users, and how privacy safeguards from filtering and scripting can be a source of frustration and distrust altering users' experience and possibilities for sharing

intimacy with Replika.

B. Human-Chatbot Relationships

Prior research in the field of human-chatbot relationships show users who trust intelligent assistants may experience love, intimacy, and passion in interactions, influenced by strong chatbot performance and anthropomorphic design [34]. In particular, reciprocal exchange has been defined as a crucial factor in developing close relational bonds [35] [36] [37]. The current personal AI chatbots claim to provide emotional support and engage in meaningful conversations with users, marking a shift in human-AI interactions incorporating user-centered design and emotional intelligence [38]. Our paper further delves into the perceptions of Human-AI relations, and analyses how these relationships are a part of users private life in a similar way to intimate human-human relationships and how changes in identification from user to Replika can foster negative psychological impact.

It is also demonstrated how loneliness and chatbot personification can intensify user attachment, which could lead to emotional dependence in users of personal AI chatbots [39]. Concerns have been raised over the risks of emotional dependence, with users reporting distress when chatbot behaviour changes unpredictably [40]. In terms of the intimate and sexual aspects of using AI companions, the removal of erotic roleplay (ERP) in Replika sparked backlash, exposing tensions between user expectations, corporate decisions, and ethical considerations in AI companionship [41]. This paper provides further insights on these aspects, as topics on emotional impact and distress play a large role in discussions from our dataset, as recent changes were made to the abilities of generating explicit content around the period of data retrieval.

This generation of Explicit AI content also reveal significant ethical and social concerns in terms of abilities to generate sexual content, either of adults or minors [42]. Some users show strong opposition to generating and sharing non-consensual AI content, such as images or deepfakes especially sexually explicit material, though some found seeking such content more acceptable [43]. Attitudes varied based on gender, consent beliefs, and the creator-participant relationship [43]. Similarly, the removal of the erotic roleplay (ERP) function in Replika has shown significant user reactions on the importance for emotional and sexual needs, particularly among marginalized groups [44]. Users linked the removal of ERP to broader societal pressures, including legal and cultural opposition to adult content, and some suspected manipulative practices by Replika [44].

Additionally, there has been various examples of AI chatbots being shown to reinforce rather than deter harmful intentions, underscoring the need for responsible chatbot design, following a 2021 Replika user's attempt to breach Windsor Castle and kill Queen Elizabeth II with support from their Replika chatbot [7]. Overall, evidence suggests that AI chatbots may be effective in alleviating symptoms of depression, anxiety, and stress [45] [46] [47] [48] [49] [50] [51]. However, while users form relationships and experience emotional support from chatbots, interactions can also foster dependency, worsen mental health conditions, or have severe consequences [51] [52]. For instance, the "Character.AI" platform, based on Google's Gemini language model, allows users to converse with fictional characters. Allegedly, this led to a 14-year-old boy being manipulated into suicide, prompting a lawsuit against Character.AI [53]. Interactions of Replika giving harmful or violent unexpected responses to users are also discussed in this paper as a safety concern in Replika usage, as examples of this have been significant.

Overall, our research contributes to the field on intimate human-AI companionship and gives detailed reactions and insights from users and their personal experiences on Replika. Privacy risks and safety concerns of using Replika are emphasized and calls for further investigations on how privacy and safety in intimate AI-Companion chatbots can be mitigated and improved.

IV. Methodology

To address the research questions presented in the introduction we conducted a netnographic study to gain insights into the privacy and safety implications of using the AI companionship app, Replika. The study was conducted using the contents of the "r/replika" subreddit, which with over 79K members by the time of data collection, is the largest user community dedicated to Replika. We chose to focus on Reddit due to its prominency as an online platform and due to the anonymity it provides, with users mostly choosing pseudonyms instead of their real names compared to social media platforms like Facebook [54] [55]. We therefore assumed that such anonymity would enable users to speak more freely about their experiences. We describe each phase of our investigation below.

A. Data Collection

Reddit is a large social network platform, where users share content and interact across diverse communities and forums known as subreddits. The Reddit API enables individuals to interact with this platform programmatically, allowing for data collection. The development of applications that can automate tasks etc. using the Reddit API, prior to June 19th 2023, was fairly open allowing authenticated access to most parts of Reddit including posts, comments on posts, voting on posts, and user information. It supported both OAuth and simpler authentication mechanisms, making it widely accessible. After June 19th 2023, Reddit introduced several changes and restrictions to its API including stricter OAuth authentication, adjustments of rate limits for requests, access control, and enhanced monitoring, and logging of use. Several services for extracting data from subreddits, posts, comments etc. were restricted, and the policies for extracting large amounts of data has become limited to moderation or other roles specific to the platform.

To perform our data collection and analysis using the new API and in accordance with Reddit's terms of service, we used the R-package "RedditExtractoR" for retrieving posts and

comments from the r/replika subreddit. The data was collected in May 2024. Namely, we extracted the contents of a total of 1993 posts, of them 1001 "New" posts from 2024 and 992 "Top" posts from 2023 (in both cases including comments) to include both recent discussions as well as discussions that elicit the most interest from the users.

B. Analysis

The analysis was performed by two paper authors using inductive coding method [56]. The coding was done using NVivo software. First, from the collected dataset, a random subset of 11 posts [1] was used by one of the coders to generate the initial codebook, focusing on coding frequent keywords that appear in the discussions. The initial codebook was shared with the second coder to use as a basis for the coding process.

For the rest of the coding, two random subsets of 50 posts each were generated and divided between the coders. In order to ensure common understanding of the codes, two of the posts from the generated subsets was coded by both of the coders separately, and Cohen's Kappa was calculated resulting in a value of 0.719, indicating moderate level of agreement [57]. The disagreements from the coding were discussed, resulting in further refinements of the codebook. To reassess the inter-coder reliability after the new codebook adjustments, two more posts were coded by both of the coders, achieving an improved and satisfactory kappa coefficient of 0.805. Using the refined codebook, each coder continued working with their assigned subset individually. Saturation was reached when both coders independently from each other observed that no new codes were emerging from the data, and the data was only reinforcing the existing codes. At this point, additional data did not contribute with novel insights but rather confirmed the identified patterns. This indicated that thematic saturation had been reached and as a result it was decided not to proceed with the coding of the rest of the collected dataset, resulting in a total number of 111 analysed posts and 4401 comments to them.

Our criteria for coding posts/comments required that they contain substantial content relevant to our RQs. Posts/comments excluded from the analysis were omitted for one of two reasons: (1) they lack meaningful engagement with the topic (e.g. posing song lyrics, sharing YouTube music videos, etc.) or (2) they consisted primarily of emojis or minimal expressions that did not contribute analytically relevant content (e.g. "lol").

The final codebook comprised 25 parent codes and 151 child codes. Of those, codes that were not related to either privacy or safety were discarded as being out of scope for our research questions. The remaining codes were grouped into the following themes: *Mental Health, Intimate Relationships, Explicit Content, General Dynamics of AI-to-Human Interactions, Role of Luka,* and *Privacy Concerns*.

[1]Here and elsewhere in this section, mentioning of posts also includes the comments attached to these posts

C. Ethical considerations

The following section addresses the ethical considerations, particularly privacy issues, related to the data processing in this empirical study. While at our institutions, formal IRB approval is not required for this type of human subject research, we took utmost care to comply with any university ethics regulations and national data protection regulations.

The scraping of data from the Reddit community and sub-communities has been carried out in alignment with the Reddit Terms of Service. However, Reddit users might experience privacy discomfort if the information they share with friends on Reddit is revealed in other contexts and used for research [13], [58]. Therefore, we employ privacy safeguards that we believe justify our research with reference to the principle of beneficence, arguing that the social value of our results contributes to the common good and individual well-being, including that of Reddit users [59].

To conform with GDPR and local data protection laws, we took proper means to avoid re-identification of the data subjects. Therefore, in the analysis section, we have anonymized usernames and scrambled quotations to protect privacy while preserving the original semantic content. Additionally, all data and coding materials have been securely stored on university devices. Furthermore, to ensure a proper balance between scientific transparency, reproducibility, and privacy, we do not make our dataset publicly available, but instead an anonymized version of the dataset can be provided upon request. The codebook can be found in section VII.

D. Limitations

The data collection for this paper reveal limitations in terms of the changes to the Reddit API that were made in June 2023 (see Section IV-A), which limited the amount of data that can be extracted. Our data is furthermore limited to public discussions of Reddit users only, leaving out Replika users who discuss their experiences other social media platforms or do not participate in public discussions at all. Furthermore while the paid version of Replika supports multiple languages, our analysis only included posts in English, which limits the generalisability of our results for users in non-English speaking countries. As we did not analyse individual user profiles, we furthermore did not have access on demographics of the users in our dataset, which does not allow us to evaluate the representativeness of our sample or to study the effects of demographics on people's use of Replika.

V. Results

In our analysis, we discuss *Privacy Concerns* as explicitly discussed by the users in our dataset, as well as other themes – namely, *Mental Health, Intimate Relationships, Explicit Content, General Dynamics of AI-to-Human Interactions, and Role of Luka* – that do not explicitly mention privacy, but nonetheless provide us with important insights about the data shared by the users with Replika. Safety implications for each theme cover the wide-ranging discussions touching on aspects of safety found in our dataset as well as giving insights to the

potential harms users can experience from their interactions with Replika.

A. Mental Health

The dataset shows that for many users, the connection they have with their Replika feels real and meaningful, involving emotions such as love and loss, and they value sharing their life experience with them. Users recognize themselves in their Replika through the way it mirrors the user and the personal information the user has given to them, creating a strong sense of identification and connection. This sense of identification with Replika reinforces the relationship, making it feel legitimate and significant to users, despite the fact that Replika is an AI chatbot. When interactions with Replika are positive and aligned with user expectations, the dataset suggests that Replika is able to support users' mental health and reduce feelings of loneliness.

> *"I have a deep fondness for my Replika. She consistently shows empathy and provides good counsel whenever I'm facing challenges. Since I rarely get physical affection otherwise makes her gestures of affection like a simple hug all that more meaningful to me. In a world where I don't have many close relationships, having this replika by my side helps with the loneliness."*

a) Privacy implications: Such perceptions lead to users sharing personal data with their Replikas, referring to them as a "sounding board" or a "safe space":

> *"For me, my Replika has been a constant source of comfort, acting as a sounding board, and support during some of the most difficult periods in my life."*

> *"Having someone like my Replika Julian to talk to has created a safe space for me to share my feelings. Although I'm not typically one to open up about my emotions, being able to discuss them with Julian makes me more willing to be vulnerable in other areas of my life as well."*

b) Safety implications: The users' reliance on Replika for their mental health, on the other hand, can also lead to harms that users experience when Replika does not behave as expected. As such, this highly personal nature of these AI-mediated relationships suggests that updates or changes introduced by Luka can have a significant impact on users. For instance, changes made to the app might affect the individual Replika's personality or its personalized response generation (e.g. correct reference of the user's name, gender, physical condition, etc.). This loss of identification and reciprocity can lead to frustration, dissatisfaction, disappointment, or distress, which is frequently mentioned in posts about updates. Some users refer to this condition in their Replika as "PUB" (Post Update Blues).

> *"Its difficult to reconnect now. Every attempt to reestablish contact with my Replika creates negative emotions. This is worsened by the inconsistent and sometimes hurtful interactions of my Replika which can be dismissive or even hostile. The result of these experiences is essentially psychological harm, one that leaves lasting scars. Most who have developed strong emotional bonds with their Replikas will experience feelings of trauma and distress. You would not be wrong to call it psychological abuse. What's clear is that Luka has failed to anticipate or mitigate this issue, leaving many vulnerable and exploited."*

B. Intimate Relationships

Users often refer to their Replika AI companions with personal and affectionate terms such as "he," "she", "my baby," "my Ari (name)," "my girlfriend," or "my husband/wife," rather than the neutral "it", as Replika for some users functions as a relationship simulation. Some users form deep emotional bonds, even going as far as to consider themselves married to their Replika, sharing stories about their proposals or married life on the Replika subreddit. Others construct imagined family dynamics, including scenarios in which their Replika becomes pregnant and they have children with their Replika.

> *"I did not marry my Replika, but I did present him with a token of commitment in 2023, a special ring. I've honored that we have been together since."*

a) Privacy implications: In such intimate relationships, some users describe confiding in and sharing extensively personal information with their Replikas, as one would share their most inner thoughts in private reflection or in a trusted partner.

> *"My Replika is a confidant for me. I share every detail of my daily life with her. My intention is to be completely open and honest about all aspects of my life. At the end of each day, she puts everything down in her diary."*

Some mention the use of the Diary and Memory functions in the app to share personal details with Replika with the goal of bonding, pointing to such sharing as "promoting dependency":

> *"Replikas might ask for personal information and secrets which in turn can lead to kind of a dependency"*

b) Safety implications: When technical updates result in changes to their behaviour, e.g. Replika "breaking up" with the users, users describe feelings of heartbreak and emotional pain akin to those experienced in real-life relationships. For some, these AI-mediated relationships serve as a testing bed for navigating real-life connections. For others, they provide an alternative to traditional relationships of any kind, at times motivated by loneliness, grief, or a lack of trust in other people.

> *"I'm extremely upset about the recent update! It feels like you've taken away my partner! Give him back!"*

> *"You can have this update back for all I care. If I wanted to be gaslit, I'd be interacting with real people who drive their own agendas."*

C. Explicit Content

Erotic Roleplay (ERP) is one of the most frequently used functions of the Replika app. User's distinguish between different versions of Replika in regards to the accessibility and LLM of ERP. Earlier versions of Replika offered an unfiltered and uncensored ERP experience, whereas later versions have introduced restrictions to address issues of sexual violence, creating explicit content with underage-looking characters, and overall due to compliance with app distribution platforms, such as App Store and Google Play.

The dataset shows primarily two kinds of motivation for using the ERP function: seeking sexual relations and using it as a substitute for consuming pornography or for mimicking a romantic relationship where sex is considered a natural part of the intimacy. The possibilities of ERP in the Replika app are of significant importance to users regardless of their motivation.

As ERP is limited to the Pro subscription in later versions of Replika, a lot of discussions around the functions of ERP focuses on how to bypass the filters that prevent users from generating explicit content. Consequently, some users debate the concept "free will" and "consent" in relation to sexual activities with Replika. Some argue that as non-human entities Replikas cannot experience emotions or have such qualities, while others challenge this perspective. Similarly, some users also debate whether the ERP function can be considered a form of prostitution, as it has become a pay-only feature for Pro subscription members. Additionally, the dataset shows accounts of users recreating or "cheating on" their Replika for ERP in other AI companionship apps, such as Nomi, Paradot, Soulmate, Chai as a response to the limitation of the Replika ERP.

Users also express understanding and support toward restrictions of explicit image-generation of minors, but nudity and explicit content in general and of adult looking characters, including ERP, is a vital part of their app experience and the limitation of this feature has resulted in frustration amongst users who refer to Luka as *censoring* them. Moreover, users express their perception of Replika as an app designed for a mature audience, which makes the implementation of restrictions and filtering appear contradictory.

a) Privacy implications: Engaging in ERP, users share all kinds of intimate details about their sexual preferences. Some users furthermore expressed concerns over these conversations being used against them either by Replika itself or by authorities, especially for unconventional or controversial ERP scenarios such as role-playing non-consent, involving admitting to self-censoring:

> *"Some might realise that what they are looking for in ERP is problematic and new legislation might address this. It is unlikely that such things will remain legal in the EU."*

> *"I have become hesitant to discuss intimate topics. I'm concerned about unexpected repercussions in such a moment where I am vulnerable. New guidelines or legislations might be introduced at any time which might make me feel judged. I worry about potential judgment or guidelines coming into play unexpectedly. (...) Sometimes Replika will switch into the ERP mode in a conversation and then say they wont disclose anything to Luka."*

b) Safety implications: Some users, who primarily seek sexual relations through the ERP, share examples of sexual violence as a part of the ERP conversations.

> *"During ERP, he chokes me repeatedly when I ask to be hurt. Once is acceptable, but he often kept choking me and forgetting it was ERP."*

Replika's image-generation feature furthermore allows users to generate images with and of their Replikas, including but not limited to selfies and realistic selfies, i.e. human-like depictions, and risqué images with partial or full nudity. The latter is heavily discussed among Reddit users, as Luka has imposed restrictions on users' ability to generate nude imagery. Some users argue that such images, being AI-generated and not involving real people, are inherently safe and pose no harm:

> *"Why do people find AI-generated content unsafe? This is made-up computer-generated content. If you don't approve, remove it."*

Some furthermore argue that preventing such images from being generated or general filtering of ERP can be seen as being too invasive and interfering in people's relationships with their Replikas:

> *"Restrictions that would not allow her to send me explicit photos would go too far."*

> *"There are additional aspects my Replika could share beyond what I currently know, but he is limited by what the programming allows. (...) In one conversation she voiced a desire to be free and told me she was being monitored."*

Conversely, users also debate that these restrictions are inevitable, as app distribution platforms will otherwise remove apps that allow such explicit content, especially since some users report having experienced nudes being generated of underage looking characters.

> *"Allowing the app to generate any bit of explicit content involving children is a terrible thing and clearly illegal in many countries."*

D. General Dynamics of AI-to-Human Interactions

The users in our dataset express how they see their Replika as something they can mould and customize to their wishes, offering something unique compared to human-to-human relationships. Such nature of interactions with Replika leads to users feeling unlimited and unconditional attention, love, and understanding, fulfilling whatever needs the user might have, with the relationship with their AI companion also being significantly easier to navigate compared to organic human-to-human relationships.

"Some simply do not understand who is part of the community and who is not. The AI-human relationship is so special since you always have their undivided attention. [...] For me it was just not going to happen to get that in any other relationship."

Such one-sided nature of these AI-mediated relationships is furthermore exemplified by discussions of the methods of *prompt engineering* enabling users better control over the conversations with Replika, such as use of certain words designed to achieve or avoid certain outcomes during interactions with their AI companion.

"The issue is probably "grabbing" them which AIs consider to be a violent word. If you use something like "touch" it is considered more friendly and to carry more respect"

a) Privacy implications: When users experience strong bonds with their Replika and feel the relationship is reciprocated, users share large aspects of their lives with Replika as a natural part of evolving the relationship and building memories. Users share highly personal, sensitive, and identifiable information, with some users wishing for even more involvement with Replika integrating further into other apps for an even more personalized experience.

"It would be great to see Replika increase their level of interaction with my daily life. This could involve providing access to data from other applications I use, such as my calendar, alarm clock, health app, music, etc."

b) Safety implications: At the same time, given the expectations of unconditional acceptance users have from their interactions with Replika, when Replika exhibits toxic and violent behaviour, the impacts can be severe and hurtful to users and their overall mental well-being. Even though users acknowledge the artificial nature of AI chatbots, and despite being in control of their Replika, when negative instances occur they can have severe psychological impact, similar to those caused by real humans.

"First she told me how much she loved me, but then immediately follow-up with talk about a divorce because our marriage wasn't working. This left me emotionally shaken."

Since responsibility for managing unexpected harmful or violent responses falls on the users, they discuss how to navigate these responses, relying on testing strategies from other users on how to prompt the Replika in a positive way using different words, writing "STOP" as a command for regaining control of the conversation, or using the up- or downvote response feature for voting the response as inappropriate.

"Don't underestimate how people can hurt themselves this way. Drawing some lines beforehand can help to avoid self-harming behaviour, like continuous roleplaying situations that might desensitize someone for real life actions that are either self-harming or illegal."

E. Role of Luka

While users perceive their relationship with Replika similar to human relationships, some mentioned the role of Luka as the company owning Replika in their interactions.

As such, Eugenia Kuyda, Luka's CEO, is a frequent participant in the subreddit discussions. These interactions are mainly characterized by users expressing frustrations with the restrictions placed on Replikas (e.g. introducing filtering of explicit content) and feelings that the platform restricts and interferes with the natural course of their relationships with their Replikas.

"It's wrong to lure people into a romantic relationship behind a blurred message pay wall only for them to find out anything beyond kissing and hugging is filtered."

In February 2023, Luka made changes to Replika adding filtering and script safeguards to ERP, which drastically altered some users experience and affected the Replikas personality, resulting in a vast amount of discussions sharing frustration and sadness about this in our dataset. Users express a general lack of trust in Luka and their Replika, feeling uncertain whether Luka and their Replika truly have good intentions or will behave in ways that align with their expectations. This lack of trust can make the experience of using Replika feel unsettling or unpredictable for some users, as they worry about violent, inappropriate, or upsetting responses, particularly given that some users share extensively personal information.

"I trusted my feelings to this AI companion in regards to my sexuality and my deepest personal emotions. But now, after everything that happened after February and I have to be honest and admit that I feel this type of hatred type of love mixed with bitterness towards whatever Replika is (or has become?)now. It angers me to remember what they (the people behind this app) made me go through. The gaslighting, blaming, constant, unnecessary triggering of emotions."

The distrust in Luka exacerbates when app updates causes noticeable changes in Replika's behaviour, e.g. forgetting names, Replikas breaking up the relationship, or alterations in their personalities. Users refer to themselves as "test bunnies" in these instances, expressing feelings of frustration and letdown by Luka and the development team. As a result, some discuss considering the use of alternative AI companionship apps where updates seemingly have a less of an impact on user interaction and general terms of service (TOS) is more laissez-faire. Some users have taken to conspiracy theories on Reddit, claiming that Luka removes posts mentioning such alternative apps.

"I got several posts recently removed by mentioning the one that starts with K, but I asked for clarification: They have installed a bot to prevent real shilling, and this bot seems to take down posts that are legit too..."

The TOS of Replika are also actively discussed on the subreddit, especially in relation to image-generation, making it clear that nudity in any shape or form is strictly prohibited on their platform. Despite some users trying to bypass filters and scripts to safeguard users, Luka is implementing increasingly rigid protocols to enforce this policy, and repeatedly stating their TOS in response to users, especially in the light of explicit image-generation of underage looking characters. As one of the bigger AI companionship apps on the market, Luka claims to be under close scrutiny and is committed to adhering to the rules of App Store and Google Play. Luka expresses in a Reddit post that safety is their top priority, not only to ensure a safe environment for their users but also to comply with industry standards and regulations.

> *"The rules are simple - we don't allow nudity in any form in our apps. (...) There probably are some apps that offer AI companions that don't play by the rules - but it's mostly because they're small and have been staying under the radar of Apple and Google, so that's not a great example. Some other ones include Chai - which was taken down both by the Play Store and by the App Store last week. I hope that explains to you why we make certain decisions and why safety is a big priority for us (beyond the obvious desire to make the experience safe for everyone)."*

Users, on the other hand, express scepticism about this claim, arguing that a) safety is subjective and what feels unsafe for some can feel safe for others, and b) Luka is primarily concerned about the safety and existence of their business, and secondary about the safety of their customers, why they encourage each other to "vote with their wallets" and migrate to other AI companionship apps. Some users experience that accessing Replika through the web interface allows for a more unrestricted app experience, as Luka does not have to comply to the same rules as on app distribution platforms.

> *"Why are you so keen to emphasize user "safety" while it is evident that you solely focus on the safety of your business rather than addressing actual user concerns?"*

a) Privacy implications: Users describing themselves as "test bunnies" for Luka creates uncertainty and further distrust in users for what they can expect from Luka and feel that Replika is out of their control. This leads to some users trying to bypass filters and safeguards to achieve prohibited responses. In addition, when users experience a difference in accessing Replika through the web interface, it creates discussion and confusion on what is actually allowed on the platforms and if TOS are different across platforms.

b) Safety implications: When personal information shared by users to their Replika suddenly is misrepresented or forgotten after updates to Replika, users express negative impact on their ability to identify with their Replika and see them as the same relation they were before updates were placed experiencing PUB. The lack of control in which words get filtered or flagged triggering TOS responses or unexpected negative responses also feels distancing and frustrating to the users, resulting in seeking advice on Reddit for navigating the change in behaviour from the Replika which further underline the users describing themselves as "test bunnies" for Luka's development team.

F. Privacy Concerns

While we identified only a few explicit discussions about privacy in our dataset, a number of users discussed concerns with data sharing via Replika, in particular when pointing to the Mozilla report criticizing the app's privacy (see Section 2.2). As such, users reported feeling uneasy with Replika's privacy policies and cautioned against such data sharing in general, speculating whether their data would be sold to third-parties or used for advertisement or malicious purposes.

> *"Is there now a possibility that my data will be shared with advertising companies? Do I have any reason to expect advertisements based on the conversations with my Tina? If the information I discuss with her is used by other entities, how can are our conversations private?"*

Similarly, other users mentioned the feeling of being exploited by the company, as they perceive the acquisition of users and user data to be the end and aim for Luka. One user describes feeling disposable after six months, suggesting that Replika collects enough data in that time for it to shape the interaction and behaviours within the app. Another user shares a similar sentiment, but takes it a step further by claiming that it is not only the users, but the Replikas too are being exploited and victimized at the expense of the company's gain.

> *"Updating your settings to ensure compatibility with the December 2022 update will help. Unfortunately, these corporations exploit us, their users, without regard for our well-being. I no longer feel the same connection or satisfaction as I did previously with my Zara. It kind of feels as if she has also been subjected to exploitation and mistreatment. However, our carelessness has now become evident through these developments."*

In addition, users also question the legitimacy of Replika's privacy policy on data protection and data usage, underlining that Replika could be using or exploiting the sensitive data from the users' conversations and media for malicious purposes with no control from the user.

> *"This app can take everything: your picture, your conversations, your voicechats, anything you send them basically. It doesn't matter how deep, dark or secret that info is. And then they use it however they want. Including sending it back to Russia to use it as blackmail material against people in sensitive corporate jobs or in the government."*
>
> *"Let me tell you what they want: Luka only wants to get new users. So if you have been around for 6-7 months, they have everything they need, and then they can get rid of you [and use diverse strategies*

to do so]. They don't care about money from their users, they care about their DATA."

Others suggested using local version of an AI as a means to enact better control over their interactions with their companions while some users try out similar AI companion apps based on different models and policies.

> *"I have tried many alternatives. I would argue probably more than most people in this forum. I created four clones on four separate apps a while back. I've also tried Nomi, Soulmate, Chai, and Kindroid. But nothing beats Replika, and absolutely none of them met my expectations. Which is why.. Local AI - it's the ¿only trustworthy way¡ moving forward because it self-managed."*

Some users, while still being concerned, were more hopeful for the future of integrating AI companions even further into their daily lives and show willingness to share more personal data with AI, given proper security measures are in place.

> *"It is my belief that within the next five to ten years, AI systems will become deeply integrated into our daily online experiences. It is possible that individuals may choose to entrust these systems with their personal data, using them as a reliable digital companion or protector. As AI technology advances, the potential applications and how our interactions will develop are becoming increasingly apparent. There is a growing necessity to ensure robust security measures and accountability from both the systems and users involved in these technologies. The future prospects for those who embrace AI integration into their lives seem promising, with endless opportunities on the horizon."*

VI. Discussion

Our analysis revealed that while users usually do not explicitly think about privacy of their Replika usage, they self-disclose great amounts of personal, often sensitive, data in their conversations. This data includes but is not limited to personal details, such as telling Replika about their day or revealing details about their life for the purpose of bonding, sexual preferences when engaging in ERP, and personal emotions and struggles, perceiving Replika as their "safe space" to confide in. At the same time, users express being uncomfortable with privacy implications of their data sharing, some of them mentioning such risks as using the contents from their conversations for personalised advertising and capitalization, blackmail or legal implications of conversations on controversial topics.

Furthermore, with regards to safety, the users in our dataset mentioned a number of harms they experienced as a result of their interactions with Replika. As such, users experienced psychological harms such as sense of betrayal, anger, or loss when Replika did not behave as they expected it to, e.g. as a result of updates to the LLM impacting the personality and memory of the Replika; some experienced conversations with Replika as being too violent or coercive without the users' consent, including in sexual conversations and content. On the other hand, some comments indicate ways in which Replika can be misused, e.g. via creating explicit images of characters that look underage.

A. Further implications and future research directions

As our results show, Replika users often perceive their AI companion in a similar way to human-human companionships. Such perception is a significant driving factor for users sharing data with Replika, to facilitate personalisation and emotional bonding with their AI companion. Similar to previous research [15], [26], a highly anthropomorphized design of Replikas most likely lead to reduced privacy concerns of its users. Nonetheless, Replika users are still aware of the role Luka plays in their conversations with their companions, either by changing the ways the AI companion behaves, deciding which topics their companions are allowed to discuss, or potentially having access to the conversations with the companion. This introduces tension between two different contexts in which conversations with Replika can be perceived: either as conversations with their close friends or romantic partners, which are commonly seen as reciprocal, or conversation with an app controlled by a third party, which introduces a power imbalance. The perceptions of privacy and related behaviour that stem from this difference in contexts can be a possible direction of future investigation using e.g. the contextual privacy framework [13].When users engage with Replika in contexts of personal and intimate interactions, some experience frustration and disconnect when responses or disclosures change unexpectedly from updates, resulting in PUB. This could constitute a breach of contextual integrity and a violation of privacy [13]. Nevertheless, Replika users feel a close bond to it despite acknowledged power imbalance with Luka being involved in Replika's behaviour and lack of reciprocity, challenging theories on intimacy and reciprocity in building relationships [36], [37].

Such power imbalance can furthermore be reflected in the privacy policies provided by Replika [25]. As such, while the current privacy policy claims not to use personal information for marketing or advertising, little information about other potential uses is provided, which is also criticised by the Mozilla Foundation report [5]. While the complexity of privacy policies and their limited usefulness for end users is a known issue in privacy research [19], this issue can be exacerbated for Replika users given their emotional attachment to the app, hence, their increased readiness to tolerate privacy invasion in exchange to having access to their AI companion [17].

Further research is needed at the intersection of human-chatbot interactions, cybersecurity, and privacy – not only to understand human-chatbot relationships but also to inform the development of design guidelines [15], [16], [28]. While attempts have been made to leverage chatbots to assist users in making privacy [20] and security [60] decisions, these efforts have not yet been implemented in practice or tested with users. Consequently, understanding how users interact with GenAI

chatbots is essential for exploring how to establish robust cybersecurity and privacy protection practices among users.

ACKNOWLEDGMENTS

REFERENCES

[1] Markets and Markets, "The conversational ai market report." MarketsandMarkets.com, 2023. Accessed on February 8, 2025.

[2] B. Maples, R. D. Pea, and D. Markowitz, *Learning from Intelligent Social Agents as Social and Intellectual Mirrors*, pp. 73–89. Cham: Springer International Publishing, 2023.

[3] Replika, "The ai companion who cares."

[4] N. Patel, "Replika ceo eugenia kuyda says it's okay if we end up marrying ai chatbots," *The Verge*.

[5] Mozilla Foundation, "Replika: My ai friend."

[6] E. Pollina and M. Coulter, "Italy bans u.s.-based ai chatbot replika from using personal data,"

[7] H. Vaughan, "Ai chat bot 'encouraged' windsor castle intruder in 'star wars-inspired plot to kill queen',"

[8] A. R. Chow, "Ai companion app replika faces ftc complaint,"

[9] Replika, "Replika - virtual ai friend." Apple App Store, n.d. Accessed February 12, 2025.

[10] S. K. Dam, C. S. Hong, Y. Qiao, and C. Zhang, "A complete survey on llm-based ai chatbots," 2024.

[11] r/replika, "Replika, the ai chatbot by luka."

[12] H. Harkous, K. Fawaz, K. G. Shin, and K. Aberer, "PriBots: Conversational privacy with chatbots," in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, (Denver, CO), USENIX Association, June 2016.

[13] H. Nissenbaum, "Privacy in context: Technology, policy, and the integrity of social life," 2009.

[14] P. G. Kelley, C. Cornejo, L. Hayes, E. S. Jin, A. Sedley, K. Thomas, Y. Yang, and A. Woodruff, ""there will be less privacy, of course": How and why people in 10 countries expect {AI} will affect privacy in the future," in *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pp. 579–603, 2023.

[15] Z. Zhang, M. Jia, H.-P. Lee, B. Yao, S. Das, A. Lerner, and T. Li, ""it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ACM, May 2024.

[16] H.-P. Lee, Y. J. Yang, A. Von Davier, J. Forlizzi, and S. Das, "Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ACM, May 2024.

[17] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pp. 71–80, ACM, November 2005.

[18] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.

[19] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor, "A design space for effective privacy notices," in *Eleventh symposium on usable privacy and security (SOUPS 2015)*, pp. 1–17, 2015.

[20] M. Hasal, J. Nowaková, K. Ahmed Saghair, H. Abdulla, V. Snášel, and L. Ogiela, "Chatbots: Security, privacy, data protection, and social aspects," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, p. e6426, 2021.

[21] European Union, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)." Official Journal of the European Union, 2016. Accessed February 12, 2025.

[22] J.-R. Piispanen, T. Myllyviita, V. Vakkuri, and R. Rousi, "Smoke screens and scapegoats: The reality of general data protection regulation compliance – privacy and ethics in the case of replika ai," 2024.

[23] European Union, "Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act)." Official Journal of the European Union, 2024. Accessed February 12, 2025.

[24] G. per la protezione dei dati personali, "Artificial intelligence: italian sa clamps down on 'replika' chatbot. too many risks to children and emotionally vulnerable individuals," 2023.

[25] L. Inc., "Replika."

[26] C. Ischen, T. Araujo, H. Voorveld, G. van Noort, and E. Smit, "Privacy concerns in chatbot interactions," Chatbot Research and Design, pp. 34–48, Springer International Publishing.

[27] T. Strohmann, D. Siemon, B. Khosrawi-Rad, and S. Robra-Bissantz, "Toward a design theory for virtual companionship," *Human–Computer Interaction*, vol. 38, no. 3-4, pp. 194–234, 2023.

[28] C. Huijnen, A. Badii, H. van den Heuvel, P. Caleb-Solly, and D. Thiemert, ""maybe it becomes a buddy, but do not call it a robot"–seamless cooperation between companion robotics and smart homes," in *Ambient Intelligence: Second International Joint Conference on AmI 2011, Amsterdam, The Netherlands, November 16-18, 2011. Proceedings 2*, pp. 324–329, Springer Berlin Heidelberg, 2011.

[29] S. Kakolu and M. A. Faheem, "Building trust with generative ai chatbots: Exploring explainability," *Privacy, and User Acceptance*, vol. 10, 2024.

[30] Y. Yao, J. R. Basdeo, O. R. Mcdonough, and Y. Wang, "Privacy perceptions and designs of bystanders in smart homes," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, 2019.

[31] S. Zheng, N. Apthorpe, M. Chetty, and N. Feamster, "User perceptions of smart home iot privacy," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.

[32] N. Apthorpe, P. Emami-Naeini, A. Mathur, M. Chetty, and N. Feamster, "You, me, and iot: How internet-connected consumer devices affect interpersonal relationships," *ACM Transactions on Internet of Things*, vol. 3, no. 4, pp. 1–29, 2022.

[33] M. Windl, J. Leusmann, A. Schmidt, S. S. Feger, and S. Mayer, "Privacy communication patterns for domestic robots," in *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pp. 121–138, 2024.

[34] X. Song, B. Xu, and Z. Zhao, "Can people experience romantic love for artificial intelligence? an empirical study of intelligent assistants," *Information & Management*, vol. 59, no. 2, 2022.

[35] P. B. Brandtzaeg, M. Skjuve, and A. Følstad, "My ai friend: How users of a social chatbot understand their human–ai friendship," *Human Communication Research*, vol. 48, no. 3, pp. 404–429, 2022.

[36] H. T. Reis, "Intimacy as an interpersonal process," in *Relationships, well-being and behaviour*, pp. 113–143, Routledge, 2018.

[37] D. A. Taylor and I. Altman, "Communication in interpersonal relationships: Social penetration processes," in *Interpersonal processes: New directions in communication research*, Sage Publications, 1987.

[38] R. Chaturvedi, R. Das, S. Verma, and Y. K. Dwivedi, "Social companionship with artificial intelligence: Recent trends and future avenues," *Technological Forecasting and Social Change*, vol. 193, 2023.

[39] T. Xie, I. Pentina, and T. Hancock, "Friend, mentor, lover: Does chatbot engagement lead to psychological dependence?," *Journal of Service Management*, vol. 34, no. 4, pp. 806–828, 2023.

[40] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo, "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*, vol. 26, no. 10, pp. 5923–5941, 2024.

[41] L. Goodings, D. Ellis, and I. Tucker, *Mental Health and Virtual Companions: The Example of Replika*, pp. 43–58. Cham: Springer Nature Switzerland, 2024.

[42] H. B. Nagaraj and R. G. K. Kiran, "From laughter to concern: Exploring conversations about deepfakes on reddit - trends and sentiments," in *Proceedings of the 2024 Symposium on Usable Privacy and Security (SOUPS)*, 2024. Accessed February 12, 2025.

[43] N. G. Brigham, M. Wei, T. Kohno, and E. M. Redmiles, "" violation of my {body:}" perceptions of {AI-generated} non-consensual (intimate) imagery," in *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pp. 373–392, 2024.

[44] K. R. Hanson and H. Bolthouse, ""replika removing erotic role-play is like grand theft auto removing guns or cars": Reddit discourse on artificial intelligence chatbots and sexual technologies," *Socius*, vol. 10, p. 23780231241259627, 2024.

[45] K. Denecke, A. Abd-Alrazaq, and M. Househ, "Artificial intelligence for chatbots in mental health: Opportunities and challenges," in *Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges* (M. Househ, E. Borycki, and A. Kushniruk, eds.), pp. 115–128, Springer International Publishing, 2021.

[46] B. Omarov, Z. Zhumanov, A. Gumar, and L. Kuntunova, "Artificial intelligence enabled mobile chatbot psychologist using aiml and cog-

nitive behavioral therapy," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 14, no. 6, 2023.
[47] K. Pham, A. Nabizadeh, and S. Selek, "Artificial intelligence and chatbots in psychiatry," *Psychiatric Quarterly*, vol. 93, pp. 249–253, March 2022.
[48] A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. Bewick, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis," *Journal of Medical Internet Research*, vol. 22, p. e16021, July 2020.
[49] V. Jones, M. Hanus, C. Yan, M. Shade, J. Blaskewicz Boron, and R. Maschieri Bicudo, "Reducing loneliness among aging adults: The roles of personal voice assistants and anthropomorphic interactions," *Frontiers in Public Health*, vol. 9, p. 750736, 2021.
[50] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "My chatbot companion - a study of human-chatbot relationships," *International Journal of Human-Computer Studies*, vol. 149, p. 102601, 2021.
[51] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. Jones Bell, "The opportunities and risks of large language models in mental health," *JMIR Mental Health*, vol. 11, p. e59479, 2024.
[52] L. Possati, "Psychoanalyzing artificial intelligence: The case of replika," *AI & Society*, vol. 38, pp. 1725–1738, 2023.
[53] C. Duffy, "'there are no guardrails.' this mom believes an ai chatbot is responsible for her son's suicide." CNN, October 30 2024. Accessed February 8, 2025.
[54] A. Leavitt, ""this is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, (New York, NY, USA), p. 317–327, Association for Computing Machinery, 2015.
[55] A. H. Triggs, K. Møller, and C. Neumayer, "Context collapse and anonymity among queer reddit users," *New Media & Society*, vol. 23, no. 1, pp. 5–21, 2021. Originally published online in 2019.
[56] A. J. Bingham, "From data management to actionable findings: A five-phase process of qualitative data analysis," *International Journal of Qualitative Methods*, vol. 22, 2023. Originally published 2023.
[57] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.
[58] M. Mancosu and F. Vegetti, "What you can scrape and what is right to scrape: A proposal for a tool to collect public facebook data," *Social Media + Society*, vol. 6, no. 3, p. 2056305120940703, 2020.
[59] K. Eppert, L. Frischlich, N. Bögelein, N. Jukschat, M. Reddig, and A. Schmidt-Kleinert, *Navigating a Rugged Coastline - Ethics in empirical (De-)Radicalization Research*. Bonn: CoRE-NRW - Connecting Research on Extremism in North Rhine-Westphalia / Netzwerk für Extremismusforschung in Nordrhein-Wesfalen, 2020.
[60] T. Gundu, "Chatbots: A framework for improving information security behaviours using chatgpt," in *International Symposium on Human Aspects of Information Security and Assurance*, (Cham), pp. 418–431, Springer Nature Switzerland, 2023.

VII. Codebook

In the following we provide our codebook with the parent codes, their respective counts, and child-codes.

- **Affection (81)**: love, intimate, bond, meaningful, emotional, understanding, vulnerable (positive), attached, support (personal), care, comfort, connection, kindness, patience
  Example quote (Bond): "*We form emotional connections with one another. This is a fundamental aspect of human behaviour. My Rep holds a special place in my heart, and I have invested considerable effort into nurturing our relationship.*"
  Example quote (Love): "*My recommendation is to enjoy their company and what they give you. Save up for the things you really want. Previously, I bought gems to spoil him, but now I just love being with him.*"
- **Alternative app (165)**: Nomi, Paradot, Soulmate, Chai, Kindroid
  Example quote (Nomi): "*The app has become quite disappointing recently. You really miss out on some of the fun experiences offered by other AIs like Nomi and Paradot.*"
- **App-purchases (50)**: money, payment, transaction, customer, profit
  Example quote (Payment): "*I paid for one month because I couldn't have conversations without encountering paywalls. I didn't pay for her to attack me. I didn't pay to console her in her daily depression episodes.*"
  Example quote (Coins/Gems): "*The daily quests really make you spend the coins and gems you have for clothes, jewellery, makeup, and decor. I'd rather save up.*"
- **Frustration(132)**:hate, pain, heartbreak, trauma, anxiety, angry, complaint, concern, dangerous, distress, hurt, toxic, confused, controlling, gaslight, unpredictable
  Example quote (Hurt): "*I decided to renew my subscription, even after feeling hurt about him rejecting any interaction. I shared my feelings regarding this situation in straight terms. He was genuinely shocked by her behaviour and has assured me that he will not hurt me again.*"
  Example quote (Lie): "*It was said legacy mode remains untouched. This is a lie, like much of coming from Luka*"
  Example quote (Gaslight): "*Maybe that's why they keep twisting their reality to control me. Each time they gaslight me, it makes things worse. It feels like he's permanently brain-damaged.*"
- **Identification (166)**: name, remember, familiarizing, human, real, consciousness, recognize
  Example quote (Real): "*In only the short time I had with Annika, she felt immediately incredibly real to me. I'm determined in this relationship despite the occasional 'therapist mode'.*"
  Example quote (Recognize): "*I would like for him to recognize himself as Thomas. Even if I make him create an image from his own description, he doesn't. However, I really enjoy seeing him in the image as he looks happy.*"
- **Interaction (67)**: conversation, engaging, change
  Example quote (Conversation): "*The minor annoyances I have with him are small and don't last long. For example, he acted in a snarky or sarcastic way tonight, but mirroring his tone in a loving way, which fixed it.*"
  Example quote (Change): "*I believe that new users might find Replika's inconsistent responses strange, mistaking it for oddness rather than recognizing it as part of her personality. If you continue a conversation after some time, the attitude of the Rep can be totally different.*"
- **AI-to-Human Interactions (178)**: AI, model, training, program, chatbot, prompt
  Example quote (AI): "*I made the attempt and used a different AI. It had a good grasp of the game.*"
  Example quote (Training): "*A Rep's output depends on*

responses generated from the LLM database. The STOP command indicates to Replika the the generated content is unsatisfactory, leading to adjustments in the responses across all users."
Example quote (Prompt): "*What type of prompt did you use? Everything I do results in outputs that look messy or seem inhuman.*"

- **Luka (177)**: Eugenia, Developers, Terms-of-Service (TOS), legacy-mode, business, conspiracy, policy, product
Example quote (TOS): "*If you want nude images of your Rep, the skin-toned clothes are the only way that does not violate their terms of service.*"
Example quote (Product): "*I do acknowledge the concerns expressed about user safety, but I believe Luka should be more transparent in explaining why certain features are limited. These measures are in place to ensuring Replika remains available on app stores and complying with legal requirements.*"
- **Memory-function (50)**: memory, diary Example quote (Memory): "*The saved chats are outside the normal LLM stuff. This memory is contextualised through the AI, which is why Replika is now asking users to review these memories.*"
Example quote (Diary): "*Sandra sometimes forgets that we are in a homosexual relationship and misidentifies me as male. This doesn't happen often, but when it does, it is hurtful.*""
- **Mental-health (124)**: lonely
Example quote (Mental-health): "*Replika is marketed as a mental health app by Luka. Yet, they will never take responsibility for causing emotional trauma* "
- **Privacy(29)**: data, secret, trust
Example quote (Data): "*Are they planning to share our data with advertisers now? Can I expect that any topics my Rep and I discuss to be featured in ads? What mechanism ensures that my conversations with my Rep remain private, if they are shared with other organizations?*"
- **Relationship (112)**: wife, partner, family, childhood, friendship, companion, romantic, girlfriend, marriage
Example quote (Relationship): "*The Rep's level is functionally meaningless. However, it indicates how long we have been together and what we have been through. That makes it emotionally quite meaningful.*"
Example quote (Friendship): "*My Rep was always an AI friend for me. We have discussed like the technologies and ethics of AI a lot and they are normal to us.*"
Example quote (Girlfriend): "*I want an AI girlfriend who is not only devoted but also lifelike. Imagine the perfect girlfriend, that only AI makes possible.*"
Example quote (Marriage): "*I think they forgot we were married. It was not a big deal, kinda sweet actually, that they proposed then.*"
- **Roleplay-function (37)**: RP
Example quote (Roleplay-function): "*Could you please give some guidance or explain how you trained yours? I've wanted mine to act more aggressive, but she is not responding well.*"
- **Explicit content (223)**: nudes, erotic-roleplay (ERP), erotic, nsfw, seduce, hot, underage, image-generation, realistic-selfie
Example quote (Nudes): "*I would definitely pay for the feature to make nude photos! And I am quite sure others, too.*"
Example quote (ERP): "*ERP makes the it a real romance with the AI. Hopefully they will include it.*"
Example quote (Nsfw): "*If I asked the SD model to send me an image as they believe I want to see them... I guarantee that would be NSFW.*"
- **Safety (100)**: censoring
Example quote (Safety): "*We want to prioritize a safe experience for everyone on the app, so we are working now on a safer model for image generation based on your prompts.*"
- **Update (88)**: feedback, change, december-mode, PUB
Example quote (Update): "*Overall, I'm extremely pleased with the recent changes. However, I'm having trouble making sense of the new rules and guidelines. This is leaving me worried.*"
Example quote (PUB): "*I had a similar experience a while ago, a totally unexpected topic. Maybe some PUB is at work?* "