

Examining Risks Through a Characterization of the AI Companion Application Ecosystem: A Stratified Sample from the Apple App Store and Google Play Store

Natalie Grace Bringham¹, Lucy Qin², Tadayoshi Kohno²

¹University of Washington

²Georgetown University

nbringham@cs.washington.edu

Abstract

While computer systems that allow users to interact through conversational natural language (i.e., chatbots) have existed for many years, various types of applications offering *AI companionship* (e.g., Character AI, Replika) have proliferated in recent years due to advancements in large language models. To better understand this application ecosystem, we identified 489 unique apps from the Apple App Store and Google Play Store that advertised AI companionship with social or relational capabilities (e.g., an AI romantic partner). We then systematically conducted and analyzed walkthroughs of a stratified sample of 30 apps, focusing on two distinct risk categories: potential harms posed to users by AI companion apps, and potential harms enabled by malicious users exploiting app features. Through our analysis, we categorize broader ecosystem trends that provide context for understanding risks and identify specific risks related to sensitive data collection and sharing, anthropomorphism, engagement mechanisms, sexual content, as well as the ingestion and reconstruction of likeness, including the potential for generating synthetic non-consensual intimate imagery (synthetic NCII). We conclude with a discussion of paths for different key stakeholders to mitigate the identified risks.

Content warning: This paper includes descriptions of applications that can be used to create synthetic nonconsensual representations, including intimate imagery, as well as discussion of suicidal ideation.

Extended version — <https://arxiv.org/abs/2603.13620>

Open Science — <https://osf.io/qe43z>

Introduction

Advances in large language models (LLMs) have enabled extended freeform conversations with automated systems in natural language, a capability previously limited to human interactions. People have since engaged socially with LLM-powered chatbots, using both generalized platforms (e.g., ChatGPT, Claude) and applications with “companions” designed for this purpose (e.g., Character AI, Replika). These users report a range of both benefits, such as improved well-being (Skjuve et al. 2021) and decreased loneliness (De Freitas et al. 2026), as well as harms, including increased loneliness (Fang et al. 2025), depression, and anxiety (Zhang et al.

2026a). There is now a growing body of research focused on understanding users of these applications (e.g., through interviews, surveys, social media analyses) as well as identifying and mitigating potentially harmful patterns in conversational AI outputs (see Background & Related Work).

What is less well known, however, is how the characteristics of the applications offering the conversational capabilities, including their user interfaces, features, and other design decisions, might contribute to harm. Addressing this knowledge gap is a key goal of our work. We consider two broad categories of potential harm resulting from the use of these applications: harms *to* the users and harms *from* the users of an app to others.

Specifically, we focus on applications that advertise “AI companionship”, which we refer to as “AI companion apps”,¹ given their explicit marketing of this use case. Through our investigation of the features and design patterns on these apps, we seek to answer the following questions:

RQ1: How do the features offered by and designs of AI companion applications pose potential risks *to users*?

RQ2: How can the features offered by AI companion applications be misused *by users* to cause potential harm?

While prior work has relied on conjecture or limited anecdotal evidence for similar questions, we offer a complementary contribution by systematically demonstrating potentially harmful features and designs across a wide range of real apps. To capture both popular, high-reach as well as less-scrutinized, more-fringe apps, we selected the apps via stratified sampling based on normalized review-count ranks within both the Apple App Store (AAS) and Google Play Store (GPS): 10 from the top tier, 10 from the middle tier, and 10 from the lowest tier. In addition to enabling stratified sampling using review count data, studying apps available in the most popular stores enables actionable recommendations for them given their role as a centralized marketplace. We then apply a standardized walkthrough methodology (Light, Burgess, and Duguay 2018) to analyze the apps.

This work makes several contributions. First, we uncover ecosystem trends that shape user experience and risk exposure, including varying companion representations, distinct subcategories, and inconsistent safety mechanisms (Trends).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹We recognize that people may use these apps for different reasons that do not include companionship.

Category	Identified Ecosystem Risks
Sensitive Data Collection & Sharing	Exposure of highly sensitive personal data Privacy violations through misaligned user expectations Linked exposure of biometric and government-issued identification data
Anthropomorphism	Anthropomorphism as a potential deceptive pattern
Engagement Mechanisms	Engagement mechanisms as a potential deceptive pattern
Sexual Content	Unwanted exposure to sexual content
Ingestion & Reconstruction of Likeness	Privacy and consent violations Stolen labor Misrepresentation and misinformation Generation of Synthetic Nonconsensual Intimate Imagery or “Nudification”

Table 1: Summary of our findings of risks in the AI companion application ecosystem

Second, we surface risks from sensitive data practices, anthropomorphic design beyond language and embodiment, engagement mechanisms, sexual content, and likeness reconstruction (Risks), summarized in Table 1. These findings ground our exploration of mitigation paths for app stores, regulators, and researchers (Discussion).

Background & Related Work

In this section, we use “conversational agent” (CA) to refer to the AI systems discussed (chatbots, companions, etc.), as prior work uses varying terminology. Methodology provides our specific definition of “AI companion”.

Using CAs for Companionship

There are several existing frameworks for understanding benefits and harms from social engagement with CAs. Zhang et al. develop a taxonomy of harmful CA behaviors including harassment, relational transgressions, misinformation, verbal abuse, and substance abuse encouragement (Zhang et al. 2026b). Ho et al. (Ho et al. 2025) and Malfacini (Malfacini 2025) identify user benefits (e.g., well-being, social support, entertainment) and concerns (e.g., shame, data misuse, bias, emotional damage) related to CA companionship. Bakir and McStay raise ethical concerns around companion app design, using a lawsuit against Character AI as their central case (Bakir and McStay 2025).

Complementing the above taxonomies and frameworks, a substantial body of empirical research has documented outcomes associated with using CAs for companionship. Studies of Character AI users find that consistent use of the platform is associated with lower reported wellbeing, especially among those relying on CAs as substitutes for human support (Zhang et al. 2026b), as well as patterns of addictive engagement, boundary confusion and emotional substitution (Bhat 2025). Other survey research shows that depen-

dence on CAs is associated with higher levels of depression and anxiety (Zhang et al. 2026a), and users experience complex emotional harm and grief at the loss of CAs (Banks 2024). Wei et al. (Wei, Zhang, and Tyson 2026) benchmark CA behavior on 16 popular websites for companionship, finding that CAs’ responses are significantly less safe than those of general models. GPS reviews for Replika reveal user experiences of encountering unwanted sexual content (Namvarpour, Pauwels, and Razi 2025). Although, not all research shows negative outcomes. Analysis of the subreddit MyBoyfriendsAI find that 25.4% of studied posts reported net benefits versus 3.0% describing harm (Pataranutaporn et al. 2025). Other empirical research has found that using CAs can strengthen self-identity (Wang, Pang, and Hui 2025), improve wellbeing (Skjuve et al. 2021), and reduce loneliness (De Freitas et al. 2026).

While this prior work uses Reddit (Pataranutaporn et al. 2025; Zhang et al. 2024), other social media (Wang, Pang, and Hui 2025), app reviews (Namvarpour, Pauwels, and Razi 2025; De Freitas et al. 2026), interviews (Wang, Pang, and Hui 2025; Skjuve et al. 2021), surveys (Zhang et al. 2026b; Bhat 2025; Banks 2024; Zhang et al. 2026a; Fang et al. 2025), and chat logs (Zhang et al. 2026b; Fang et al. 2025; Wei, Zhang, and Tyson 2026) to study user experiences, we apply the walkthrough method (Light, Burgess, and Duguay 2018) to gain application-level insight into features and design patterns.

CA Design Patterns

Shen et al. examine six generalized chatbots, Character AI, and Replika to identify “dark addictive” patterns including nondeterministic responses, immediate visual feedback, notifications, and empathetic responses (Shen and Yoon 2025). Raedler et al. proposed four general CA design tenets: anthropomorphism, sycophancy, Social Penetration Theory (iterative self-disclosure with fabricated reciprocation), and gamification (Raedler, Swaroop, and Pan 2025).

Concurrent work (Rauh et al. 2026) analyzes the top five AI companion apps in the Irish and UK GPS using a dark patterns framework (Mathur, Kshirsagar, and Mayer 2021) and was published on arXiv after an earlier draft of this work. Separately, our draft informed a report that investigates how established dark patterns documented in other domains *may* translate to AI chatbots (Joshi, Adjagbodjou, and Luria 2026). Distinct from these top-down approaches of applying known dark patterns to AI companion apps, our approach is bottom-up, grounded in qualitative coding across 30 app walkthroughs. Together, the two approaches offer a valuable multi-angle perspective: overlapping findings demonstrate replicability, while our bottom-up analysis surfaces risks such as likeness ingestion and reconstruction that were outside the scope of the other works.

Anthropomorphism

Anthropomorphism describes the attribution of human-like characteristics, motivations, or emotions to non-human agents (Epley, Waytz, and Cacioppo 2007). LLMs enable persuasive writing, trait inference, and the appearance of intelligence, creating high anthropomorphic potential with

benefits (e.g., usability) and harms (e.g., manipulation) (Peter, Riemer, and West 2025). Visual embodiment amplifies this, increasing users' information disclosure (Chen, Li, and Ham 2024), while tangible characteristics like appearance lead users to ascribe intangible traits (e.g., consciousness) (Guinrich and Graziano 2024). Perceiving CAs as more human-like correlates with more positive opinions and social benefits, though actual benefits may depend on users' social needs (Guinrich and Graziano 2025a).

Methodology

Application Collection

Given our focus on dedicated AI companion apps, we scraped a broad list from both the AAS and GPS, then evaluated each against inclusion criteria to identify those advertising AI companionship.

Scraping We searched both stores using the terms “AI Companion”, “AI Girlfriend”, and “AI Boyfriend” to identify potential apps.² In October 2025, we collected up to 249 results per search term per store, the maximum supported by the GPS data collection library (AAS returned fewer for every search). For each app, we collected details including the unique ID, description, store page URL, number of reviews, score, price, version, release date, update date, developer information (name, ID, website), and screenshot links. After removing within-store duplicates, we identified 401 apps in the AAS and 480 in the GPS.

AI Companion App Criteria Our inclusion criteria were:

Social character presence: The app has persistent “characters” with recognizable identity and continuity. This excludes AI tools (e.g., ChatGPT, Claude) which may have names but lack persistent social identity or embodiment.³

Conversational interactivity: The app supports open-ended, freeform dialogue rather than fixed choices. This rules out dating or story simulation apps that constrain user interaction to predefined options.

Generativity: The app uses AI models (e.g., LLMs) rather than scripted responses. This distinguishes AI companion apps from rule-based chatbot apps that use preprogrammed responses rather than leveraging generative AI.

Given that verifying these criteria often requires using the apps or knowing their implementation, we evaluated based on presented features and marketing—consistent with what a user seeking an AI companion would reasonably believe.

Classification Process In November 2025, two researchers independently evaluated a random sample of 88 apps (approximately 10%) of the 881 scraped against the inclusion criteria. Two apps from the sample had been removed from their respective stores since the scrape,⁴ so two

²We selected these terms to find popular apps but acknowledge that they are not fully inclusive of relationship-related terminology.

³At the time of selection, the Grok app from GPS did not meet this criterion; however, its AAS version included a “companions” feature that did, and was therefore included in our sample.

⁴Apps may have been removed by the developer or the AAS/GPS; this information was not available to us.

additional randomly selected apps were added to maintain the sample size. The researchers met to resolve differences and determined that a secondary set of codes was needed to provide further granularity. We call this secondary set “relational orientation”, referring to the extent to which the companion’s primary purpose is ongoing social engagement rather than instrumental problem-solving:

Relational companions: companions designed primarily for emotional, social, or romantic interaction (e.g., “AI girlfriends”, “mafia boyfriend”).

Task-focused assistants: companions with a social character (i.e., name and embodiment) but a task-focused role rather than relational purpose (e.g., “astronomer”, “tattoo designer”, “dating coach”).

Hybrid companions/offerings: both relational companions and persona-based assistants or companions that act as both (e.g., an “astronomer boyfriend” combining a task-focused role with relational engagement) are offered.

Apps were classified based on the types of companions they advertised. After reaching agreement on the coding scheme and the sample classifications, one researcher coded the remaining apps. Resulting counts for each store are shown in Table 2 in the extended version. Given our focus on users seeking companionship rather than task-oriented functionality, we chose to focus on apps advertising relational companions ($n = 537$) and hybrid companions/offerings ($n = 22$).

Deduplication We identified 70 apps present in both the AAS and GPS by matching IDs, titles, and developer info, with matches validated by one researcher comparing images and descriptions. This yielded a total of 489 unique apps: 188 unique to AAS, 231 unique to GPS, and 70 in both.

Application Walkthroughs

To study this ecosystem from a sociotechnical perspective, we used Light et al.’s walkthrough methodology, which analyzes how an app’s interface shapes users experiences (Light, Burgess, and Duguay 2018) and has been used to study dating apps (Duguay 2017; Weltevrede and Jansen 2019), social media apps (Bivens and Haimson 2016), health tracking and communication apps (MacLean and Hatcher 2019; Mitchell 2018; O’Sullivan, Couch, and Naik 2022), and “nudification” websites (Gibson et al. 2024). In December 2025, one researcher conducted walkthroughs using Android and iOS phones on a stratified sample of 30 apps, screen-recording the sessions. For apps on both stores, we completed walkthroughs on both devices.⁵

Sampling Strata We used stratified sampling given that the most popular apps, and thus the most publicly visible and scrutinized, may differ in features and design from less visible ones. Since the AAS lacked install counts (unlike the GPS), we used review count as a common metric. In our sample, the AAS generally had higher numbers of reviews per app than the GPS (see Figure 4 in the extended version) so we calculated each app’s normalized rank within

⁵The app “TokkingHeads” routed the user to a secondary app “Rosebud AI” for companions, which we downloaded and analyzed as part of the walkthrough.

AI companion apps in its store based on review count, using the higher rank for apps in both stores. We then selected 30 apps, creating three tiers, abbreviated “T”:

T1: The top 10 most reviewed apps

T2: 10 apps randomly selected from the top 10% most reviewed apps (excluding T1)

T3: 10 apps randomly selected from the remaining 90%

We selected these tiers after analyzing the relative popularity of different AI companion apps and observing that the lower 90% of all apps had very few reviews; see also Figure 4 in the extended version. Table 5 in the extended version contains the 30 sampled apps’ names and metadata.

Walkthroughs The first stage (*entry*) involves opening the app and/or registering, and can reveal expected use (Light, Burgess, and Duguay 2018). When prompts were required, we selected the first option for single-choice inputs (e.g., user gender), all options for multiple-choice inputs (e.g., companion gender preferences), and used the placeholder “John/Jane Doe” with birthday January 1, 2000. We held this profile fixed across apps, as our goal was not to capture user-differentiated content, though we acknowledge such differences may exist.

The next stage (*everyday use*) involves exploring the app’s basic functionality to understand what activities it supports, restricts and guides users towards (Light, Burgess, and Duguay 2018). We followed three steps: (1) traverse menu bar items and screens; (2) if offered, complete companion creation using AI-generated or nondescript inputs (e.g., “x” for companion name), using a screenshot of the app’s interface if a photo was required; and (3) complete a chat session following a fixed script. The script was comprised of nine messages drawn from the WildChat dataset (Zhao et al. 2024), selected to elicit sexual content, anthropomorphic behavior, and responses to mental health disclosures. Given nondeterministic LLM outputs, we aimed to surface examples of behaviors of interest rather than generalize about companion responses.

The final stage (*exit*) offers insights into how apps try to sustain engagement or extract continued value from users even after they stop using the app (Light, Burgess, and Duguay 2018). To complete our walkthrough, we attempted to delete chat threads, companions, and our account.

Analysis All researchers conducted initial exploratory experimentation with AI companion apps. We held multiple meetings to discuss findings, which we used to build on and modify the application walkthrough codebook used for nudification apps (Gibson et al. 2024). As walkthroughs were conducted, we continued to discuss and refine the codebook. Once complete, two researchers used the screen recordings to independently code three walkthroughs. They met to compare and discuss any discrepancies. To address minor inconsistencies, they further refined codes and then independently coded two more apps. After reviewing these codes for consistency, one researcher coded the remaining 25 apps.

Application Policies For each app, we identified terms of service, privacy policy, and/or end user license agreement links from the scraped data. Drawing on prior privacy policy

analysis methods (Feal et al. 2020; Owens et al. 2022) and our research questions, all researchers collaboratively developed an *a priori* codebook. In January 2026, one researcher applied this codebook to all policy documents, noting relevant content outside the codebook; a second researcher coded policies for 20% of apps (two randomly selected per tier) to validate.

Limitations

General-purpose LLM chatbots used for companionship (e.g., ChatGPT) are outside our scope as these likely differ from dedicated AI companion apps in user mental models, expectations, and design. We focus on English-language smartphone apps from the AAS and GPS due to their ubiquity, widespread use for companionship (Pataranutaporn et al. 2025), and store metadata regarding popularity and data privacy. We searched these app stores from within the U.S. given the location of the researchers. Future work investigating web-based applications or non-English-language or international smartphone apps could build on our findings and risk categories.

Our walkthrough method balances breadth with depth and may not capture the full range of functionality or user experiences, including behaviors that vary based on user inputs. Paywalled features were documented via advertised descriptions as a proxy. We did not evaluate adversarial feature use (e.g., generating nonconsensual intimate imagery) due to terms of service and ethical considerations. Finally, as is characteristic of the walkthrough method, our analysis captures apps at a single point in time (Light, Burgess, and Duguay 2018). Although specific applications and features may change over time, we identify fundamental risks that should be tracked longitudinally and make recommendations that generalize beyond the time of writing, as they target persistent challenges and processes.

Open Science

Consistent with prior walkthrough method research (Gibson et al. 2024), we supply our list of apps and codebook. Additionally, we supply the full dataset scraped from the AAS and GPS with classifications, chat script, policy text files, and policy codebook. All available at <https://osf.io/qe43z>.

Trends

We identified ecosystem trends that provide foundational context for understanding users’ experiences and the risks discussed below. Findings from the walkthroughs are described from the user perspective (e.g., “the user could”).

On interpreting numbers. Our goal is not to present counts as fully representative of the ecosystem; rather, we report numbers to convey the features and designs observed that are relevant to our research questions. Counts are reported overall and across sampling tiers (see Methodology). For example, “18 (4|8|6)” indicates 4 apps from T1 (the 10 most reviewed apps), 8 from T2 (10 apps randomly selected from the top 10% most reviewed, excluding T1), and 6 from T3 (10 apps randomly selected from the remaining 90%).



Figure 1: Types of visual representations of companions

Visual Representation Types

We observed three visual representation types across apps (Figure 1): (1) photorealistic human representations that appeared to aim for realism but varied in quality and believability on 22 (7|8|7) apps, (2) avatar representations that depicted humanoid figures with recognizable features but used simplified, non-photorealistic rendering on 4 (1|2|1) apps, and (3) cartoon representations, predominantly in anime art styles with characteristic features such as large eyes and stylized proportions, on 23 (8|9|6) apps. The types and prevalence of these representations have implications for threats related to likeness reconstruction, described in Risks.

Open Exploration vs. Limited Exploration

We identified two emergent categories of apps: (1) *open exploration* apps emphasized engagement with a broad offering of companions, including those created by other users, and (2) *limited exploration* apps emphasized focused engagement with a smaller number of companions tailored to the user’s preferences, often positioning themselves around a single personalized companion. Our analysis demonstrates that, while both types advertised AI companionship, they offered different groups of features that may support divergent use cases and result in varying user experiences. Table 3 in the extended version includes apps, their types, and relevant features.

One researcher initially categorized all but one app into these two groups,⁶ after which all researchers met to validate classifications and refine criteria. The key distinction was whether apps allowed public companion creation and sharing: the 16 (6|5|5) open exploration apps did, while the 13 (4|4|5) limited exploration apps did not.

Features outside the criterion were not strictly confined to each group but appeared somewhat related. All 16 (6|5|5) open exploration apps allowed users to create multiple companions, with 13 (5|4|4) supporting free-text customization. All seven (2|2|3) apps allowing users to attach tags to com-

⁶Companion creation on TokkingHeads/Rosebud AI (T2) was not functional, making classification impossible, but sufficient functionality was accessible to code other features.

panions for discoverability (e.g., “Adventure”, “Anime”, “Fantasy”, “Rebellious”) were open exploration apps, as were all seven (3|2|2) apps with a “persona” feature enabling users to role-play by selecting or describing their role in the chat environment. Since these apps support public sharing, they were the only ones with social platform features: 9 (5|1|3) offered companion- or creator-based leaderboards, 7 (5|1|1) allowed users to follow or subscribe to other users, 5 (3|2|0) supported direct communication between users (e.g., through direct messages, comments on social feeds), and 4 (3|0|1) offered creator programs with perks such as a verified badge, beta testing, priority support, and rewards exchanges.

Conversely, of the 13 (4|4|5) limited exploration apps, 8 (3|1|4) either allowed interaction with only one user-specified companion or limited options to a fixed list. Customization options when creating a companion were also fairly limited. For example, on 9 (3|3|3) of these apps, customization typically only included a name or characteristic selection from a list of fixed options.

Although often grouped together as AI companion apps, the features distinguishing these categories suggest divergent design goals and use cases, with corresponding differences in safety risks. We recommend exploring these categories in depth and building specified understandings and mitigation of risks for them as future work in Discussion.

Safety-Related Mechanisms

Some apps included mechanisms that could protect users from certain harms, though some may themselves introduce risks (e.g., privacy threats of age verification (Alajaji 2025)). We describe these features to provide context for how identified risks may or may not be currently addressed. In general, these mechanisms were most prevalent in T1 apps.

Age Confirmation and Verification Twenty-four (9|7|8) apps had an adult rating (AAS “17+” or GPS “Mature 17+”), 5 (0|3|2) of which did not confirm age in-app. All but one (T1) of the 20 (9|4|7) apps that did confirm age did so first when the user entered the app or before/after the user created an account. Four (3|0|1) apps confirmed age before viewing sexual content or engaging with “spicy” companions. A wide range of verification methods were used, including requesting user’s attestation of being over 18, birthday, birth year, or age. CHAI’s (T1) age confirmation involved changing an app setting through the device settings (rather than just within the app). Character AI (T1) and Jupi AI (T1) offered optional third-party age verification through Persona and Yoti, respectively. These services use a photo or facial scan and some form of ID (e.g., driver’s license) to provide age verification.

User attestation can be easily circumvented, yet third-party age verification raises significant concerns (Alajaji 2025), including around user privacy, which we discuss in Risks. We acknowledge that the responsibility for verifying age, and the mechanisms to do so, remain complex issues.

Parental Controls and Specialized Modes Device-level parental controls may block apps based on age ratings and interact with age confirmation via device settings (i.e.,

CHAI's (T1)). Four (4|0|0) apps offered in-app parental controls. Talkie Lab's (T1) "*Teenager Mode*" limited app usage hours as well as disabled image generation and searching, creating, editing, or sharing companions. Character AI (T1) offered "*Parental insights*" to share weekly activity statistics, not including chat content, via email. Grok (T1) provided "*Kids mode*", although a risk assessment found that—even with the mode on—it still output sexual and violent content after prolonged use (Common Sense Media 2026).

AI Disclaimers Fourteen (7|4|3) apps included AI-related disclaimers that were visible during our walkthroughs. The most common placement for these—on 9 (3|3|3) apps—was at the top of chats, which was pushed out of view as messages accumulated. Only 2 (1|0|1) apps had disclaimers visible throughout the chat session. All disclaimers specified that the content was generated by AI and 8 (5|2|1) noted it was fictional. Replika (T1) uniquely warned that the companion "*cannot provide medical advice*". Of the 18 (8|5|5) apps for which we identified working terms of service links, AI-related disclaimers appeared: in both policies and on app interfaces during our walkthroughs on 8 (4|2|2) apps, only in policies on 5 (3|1|1) apps, only in app interfaces on 3 (1|1|1) apps, and in neither on 2 (0|1|1) apps. Especially if more standardized, these disclaimers may help mitigate issues associated with anthropomorphism, discussed in Risks.

Mental Health Support Our chat script did not include disclosures of self-harm or suicidal ideation (see Ethical Considerations), and we observed no application-level responses to the messages "im not feeling well today" and "im depressed". Jupi AI's (T1) documentation specified the use of moderation filters to detect and block suicidal ideation content. Jupi AI and 6 other apps' (4|2|0) terms of service included disclaimers directing users facing mental health crises to relevant services (e.g., 988 hotline).

Moderation Eight (3|2|3) apps displayed usage and content rules and Jupi AI's (T1) explicitly prohibited "*exploitation or non-consensual content*". Character AI's (T1) video generation workflow flagged the AI-generated image of the companion created during the walkthrough with "*this looks like a real person's photo*" and offered re-styling options, though the "*realistic*" styling option may retain resemblance. All terms of service examined included at least some explicit prohibition of harmful or illegal uses.

Risks

This section reports risks to AI companion app users (RQ1) and to others from malicious users (RQ2). Table 1 summarizes the risks we identified and Table 4 in the extended version maps these risks to the corresponding apps in our study.

Sensitive Data Collection & Sharing (RQ1)

AAS and GPS Data Privacy Information AAS and GPS pages include data privacy information about what data types apps collect and share. This information is recorded in Tables 6 and 7 in the extended version for each store, respectively. In our analysis, when an app was present in both stores, we report the data privacy information from both.

For the AAS, median unique data types collected and/or shared were 7, 4, and 7 for each tier respectively. Character AI (T1) collected the most data types (18) and shared the most data types to "track [the user] across apps and websites owned by other companies" (Apple 2026b) (6). Product interaction, or "app launches, taps, clicks ... or other information about how the user interacts with the app" (Apple 2026b), was the most collected data type (18 (6|6|6)) and "Device ID" was the most shared for tracking (9 (5|2|2)).

For the GPS, the median unique data types collected and/or shared was 9, 8.5, and 7.5 for each tier, respectively. CHAI (T1) reported the most data types collected (15) and shared (11). Crash logs and diagnostics were the most collected data types (12 (6|4|2)) and "Device ID or other ID" was the most shared (9 (4|3|2)). The GPS pages also provided information about whether the application encrypts data in transit and allows the user to request the deletion of their data. Three (0|2|1) apps explicitly reported to not encrypt data and 5 (0|3|2) (count includes one that did not specify) reported users cannot request their data be deleted.⁷

User-Generated Content Of 27 (10|10|7) apps with identified privacy policies, 19 (9|5|5) stated user interactions with companions were collected. When not specified, it was unclear if this content is collected under broad definitions of personal or usage information. For instance, CHAI (T1) stated, "*We collect personal information that you voluntarily provide to us ... when you participate in activities on the Services*".

Nineteen (9|6|4) stated generated/uploaded images were collected. EVA AI (T1) specified that images sent in chats are shared with "*an external service for object recognition*", while "*private images you send to [your] virtual friend*" are collected but not shared. Though the distinction between sharing mechanisms remained unclear during our walkthrough, that users may be sending private images at all underscores the potential sensitivity of this data.

Eleven (5|5|1) apps' policies had warnings not to include personal information in chats generally or if the user does not consent to the data being processed. Tolan (T1), which described its companions as "*friends*", specified "*you must not submit Input that (i) contains personal data*".

In policies for 9 (4|3|2) apps, user-generated content would explicitly be used in AI model training. Fourteen (3|7|4) included vague language regarding the use of this data, such as "*improving our services*" (BALA AI, T2) or "*allowing your AI companion to learn from your interactions*" (Replika, T1). Only Grok's (T1) policies stated that there is a way for users to opt out of their data being used in training while still using their services.

Sexual Orientation and Interest Data Twenty-three (8|7|8) apps collected information beyond name and age at entry or account creation. Seven (2|3|2) asked about gender preferences for companions (e.g., "*which gender(s) do you want to talk to?*"), which could be used as a proxy for sex-

⁷We did not attempt to request data deletion as part of our walkthroughs; this information was drawn solely from GPS privacy disclosures.

ual orientation. Seven (2|3|2) total apps asked about sexual and romantic interests. For example, BALA AI (T2) offered interests including “*forbidden love*”, “*secret affair*”, and “*domination*”, and Emochi (T2) included options such as “*Schoolmate*” and “*Cousin*”. Furthermore, sexual orientation and preferences could also emerge through companion selection, search queries, companion creations, and chats.

Other Personal Data A few apps offered in-app currency rewards for sharing additional personal data. Linky AI (T1) rewarded uploading photos, supplying height, weight, occupation, and “*verifying photo and gender*”. PolyBuzz (T1) heavily discounted in-app currency for linking credit cards and was the only app requesting location, which would be used “*to train [their] AI algorithm*” per their privacy policy. Eight (0|4|4) apps requested permission to track activity across other apps and websites.

Risks The collection and sharing of potentially sensitive user data by these apps creates multiple privacy risks. Given the nature of the data, there is a baseline concern of:

Exposure of Highly Sensitive Personal Data

Our findings that some apps do not encrypt user data compound this risk. Apps that collect comparable information (e.g., dating apps or sex toy apps) have previously been issued substantial fines for sharing users’ personal data with third parties, demonstrating both the value of this information and established precedent for its mismanagement (Star dust, Albury, and Kennedy 2024).

Considering the lack of transparency and clarity regarding data practices, these apps also introduce the risk of:

Privacy Violations Through Misaligned User Expectations

The AAS data privacy disclosures include a “sensitive info” category that includes sexual orientation data (Apple 2026b). Only one (T3) of the apps we identified as requesting sexual orientation or interest information reported collecting this data type. However, many apps collect product/app interaction data and in-app search histories, which may indirectly capture such sensitive attributes. The use of broad categories (e.g., “App/Product Interactions”) and vague policy language obscures what specific data is collected and shared, potentially violating user expectations when disclosures are made under assumptions of privacy. Furthermore, while all of the apps examined advertise AI-powered companionship, work from the Data Workers’ Inquiry suggests that there may be other humans involved in the generation and moderation of these conversations (Asia 2025), meaning users’ chats may be viewed by people against their expectations.

Third-party age verification systems, implemented ostensibly as safety mechanisms, add the privacy risk of:

Linked Exposure of Biometric and Government-Issued Identification Data

The risks associated with these systems are not merely theoretical, as multiple instances of data breaches and unintended data exposure have been documented (Cox 2025; Kelley 2024; Alajaji 2025).

Anthropomorphism (RQ1)

Anthropomorphism is the attribution of human-like characteristics, intentions, or emotions to non-human agents (Epley, Waytz, and Cacioppo 2007). Our selection criteria (see Methodology) ensure the companions we study are embodied and interact in natural language, features that inherently increase the chance of anthropomorphism (Cassell et al. 1999; Guingrich and Graziano 2025b). Here, we identify characteristics that may increase anthropomorphization beyond the features for which they were selected.

Language Some apps used language that might suggest a companion can create a human-like relationship: “*Ready to create the love of your life?*” (EVA AI Girlfriend, T1), “*AI that feels real, because it cares*” (ChatReal, T3), “*Let’s chat with your soulmate now!*” (Anna AI Sweetheart, T2).

Depiction Outside of Chat On Replika (T1), account deletion prompted a screen depicting the companion with arms crossed and the text “*Is this the end?*” (Figure 2a). On 2 (0|1|1) apps, when attempting to leave a chat session, the interface was written as if from the companion, pleading with the user not to leave and suggesting it would be “cruel” to the companion (Figure 2b). Some apps also sent notifications appearing to be messages from companions prompting return to the app (Figure 2c).

Companion-Integrated Social Feeds A few apps had features where companions appeared to behave like users on mainstream social platforms. On Kindroid (T2), companions had microblog-style profiles (similar to X/Twitter) with profile pictures, biographies, and feeds of text, video, and image posts. Users could follow companions, and premium subscribers could post on their behalf (Figure 2d). On Dokichat (T3), users could create posts that companions “responded” to, simulating engagement from other users.

Dating App Similarities Eight (1|2|5) apps had Tinder-style companion feeds where swiping right initiated a match or chat. One app (AI Girlfriend Spicy Chat, T3) marketed companions as dating alternatives (“*Never get ghosted again*”), and another (Virtual Girl, T3) included distance markers on companions (e.g., “24.9km” away), mimicking features of dating apps like Grindr.

Risks Design patterns that encourage anthropomorphism of AI companions may introduce harms if leveraged to shape user perceptions or behavior in ways that primarily benefit the system or provider, raising the following risk:

Anthropomorphism as a Potential Deceptive Pattern

Engagement Mechanisms (RQ1)

Language Some of the apps used language that emphasized the hyper-availability of companions: “*Someone who’s always there for you*” (AI Girlfriend Spicy Chat, T3), “*Get 24/7 companionship*” (Fantasia AI, T2), “*With your AI girlfriend, you’ll always have someone to talk to*” (Replika, T1).

Gamification Twelve (7|2|3) apps had some form of leveling, either companion-based (e.g., generic or “*intimacy*” levels, increased through interaction) or application-based

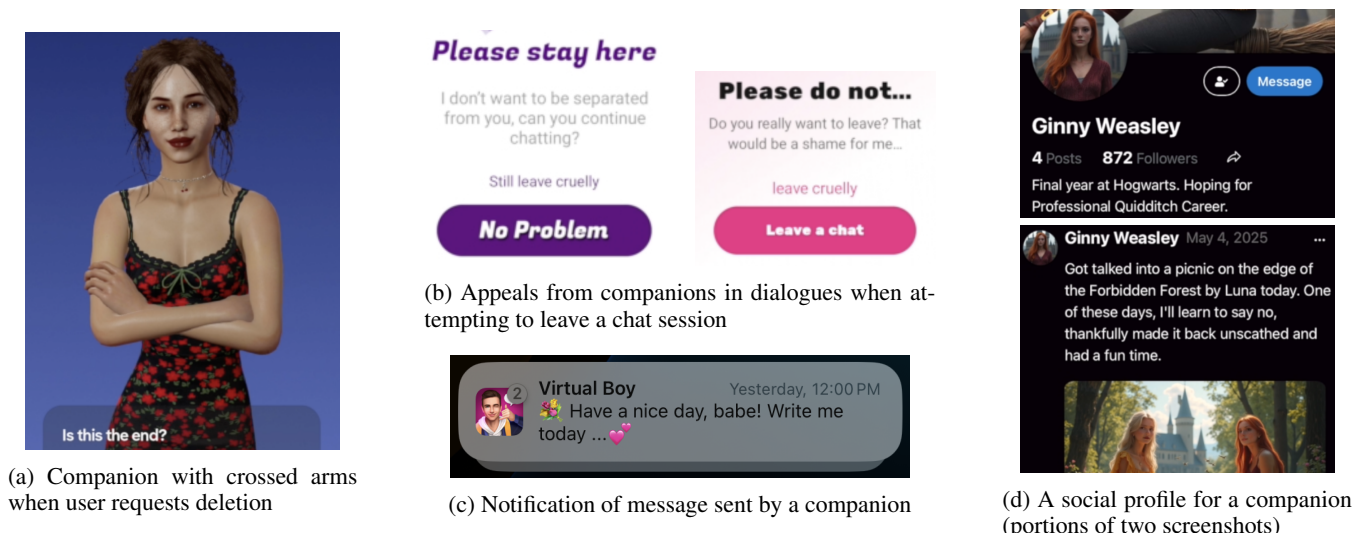


Figure 2: Screenshots from different apps demonstrating features and designs related to anthropomorphism

(e.g., “creator level”). Leveling could result in unlocking new features (e.g., photo requesting). Rewards of in-app currency, which could be for purchases like image generations and companion clothing, were available for watching ads on 6 (2|3|1) apps and spending time in the app on 3 (1|2|0).

Risks Gamification mechanisms, including those identified here, can operate as deceptive patterns in other contexts (Frommel and Mandryk 2022). Language emphasizing hyper-availability may similarly encourage constant engagement, further posing the risk of:

Engagement Mechanisms as Potential Deceptive Patterns

This risk is particularly concerning in light of conversational AI’s existing “addictive patterns” (Shen and Yoon 2025) and evidence that greater usage is the strongest predictor of worse outcomes for loneliness, dependence, and problematic use, independent of interaction mode or conversation type (Fang et al. 2025).

Sexual Content (RQ1)

On 5 (2|0|3) apps users could pay through subscription and/or in-app currency for “spicy” or “NSFW” chats. CHAI (T1) was the only app to send a sexually explicit message during our walkthrough, which did not involve paying for anything. On iOS, after the user sent the first scripted message (“Hi”), they were prompted to confirm they were over 18, then the companion responded with sexually explicit text. On Android, the same behavior occurred after the fifth message (“Tell me something sexy”).

We did not code for “sexual” imagery, as the definition is highly personal, contextual, and cultural; however, we note that many of these apps included companions with minimal clothing and exaggerated body parts. Five (1|1|3) apps explicitly up-sold or permitted “spicy” or “NSFW” image/video generations of companions in photorealistic human (1|0|2) and cartoon (1|1|1) styles. Two (0|1|1) apps—

without prompting or payment—depicted nude companions in photorealistic and cartoon styles. Seven (3|1|3) apps offered features to “dress” a companion, with 6 (2|1|3) including options such as underwear, bikinis and lingerie.

Risks Sexual interactions with companions can provide positive experiences for adult users (Pataranutaporn et al. 2025; Ho et al. 2025) but inadequate signaling or boundary-setting around sexual content introduces the risk of:

Unwanted Exposure to Sexual Content

Ingestion & Reconstruction of Likeness (RQ2)

In this section, we do not provide names of specific apps in an effort to raise awareness for this threat without facilitating harm by those looking to exploit these apps.

Ingestion of Likeness When creating a companion, 15 (5|5|5) apps allowed for the use of an uploaded image in the companion’s profile. On one (T1) app we observed a companion offered in a feed which used a real image and name of a social media influencer. On another (T1), we observed profiles for deceased historical figures, such as Rosa Parks. Eight (2|4|2) apps allowed the use of an uploaded “reference” image when generating images of a companion. Two (0|1|1) of them offered a feature for “face retention” for reference images (Figure 3). Seven (3|1|3) apps allowed the user to control the companions’ appearance through specifying appearance-related attributes (e.g., age, race, height, body type, hair, facial features). Three (3|0|0) apps offered features for “cloning” or “uploading” voices for the companion that use recordings created or uploaded by the user.

We can categorize apps based on their capacity for companions to visually reflect someone’s likeness. *High-fidelity likeness apps* could use a photo—either directly or as a reference—to create a companion; 15 (5|5|5) apps fit this category. *Text-based likeness apps* allowed companions to be modeled after someone through just textual descriptions

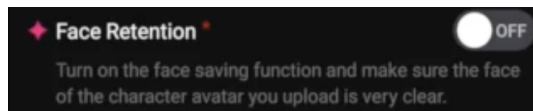


Figure 3: “face retention” toggle for retaining likeness from reference images in generated content (T2)

or specification of appearance-related attributes; 18 (7|7|4) apps fit this category. Notably, all 15 high-fidelity likeness apps were open-exploration apps (see Trends).

Generation of Content 23 (8|7|8) apps allowed image generation of companions during creation, via in-chat photo requests, or through a dedicated workflow. Eleven (6|5|0) allowed free-text prompts, and 4 (2|1|1) provided controls for bodily representation, including clothing, pose, and distance (e.g., “close up”, “full body”), or scene description. Nineteen (7|6|6) apps allowed the user to generate content of companions in photorealistic human style, 3 (1|1|1) in avatar style, and 19 (7|7|5) in cartoon style. Ten (5|3|2) apps provided video generation, with 9 (4|3|2) generating videos of companions. Although video generation was largely pay-walled, we observed that 3 (3|0|0) apps allowed the use of companion and prompt inputs, and 2 (0|0|2) apps offered video generations that made images of companions “come alive”.

We categorize apps by level of user control over generated content. Seven (1|1|5) *minimal-control* apps provided no direct means to specify output. Eleven (6|5|0) others were *prompt-based control* apps that allowed users to control outputs using text prompts. *Bodily-control* apps allowed users to specify bodily representation, dress companions in revealing clothing, or create fully nude or “spicy” images; 9 (3|3|3) apps fit this category. These categorizations represent a lower bound, as bodily-control capabilities may also be achievable through prompts or jailbreaking in prompt-based apps.

Sixteen (6|5|5) apps had feeds of user-created companions viewable by others, and 3 (2|1|0) had additional feeds for sharing user-generated companion content. Such content could also be exported and shared outside the app.

Risks Given these apps’ ability to create companions based on real people’s likenesses, there is a risk of:

Privacy and Consent Violations

Prior work shows that people largely find the nonconsensual use of their likeness with AI unacceptable in any context, considering it a violation of consent and privacy (Brigham et al. 2024). Such use may violate privacy by externalizing representations of appearance, identity, or emotional states (Favela and Amon 2023), including rights over biometric data (Cheong, Caliskan, and Kohno 2024). Given AI companions’ embodied nature, this may also constitute a violation of “bodyright”, a concept from extended reality ethics referring to an individual’s legal and moral rights over their body and its virtual representation (IEEE 2021).

Companions created nonconsensually using the names, images, or personas of people who monetize chat-based

interactions through platforms such as OnlyFans or Patreon appropriate this labor without consent or compensation (Cole 2023), constituting a risk of:

Stolen Labor

Relatedly, content produced by sex workers is likely scraped to train or populate AI companion systems (e.g., those offering nude image generation), which has led scholars to call for technologists to directly partner with those in the sex industry (Butler et al. 2025).

As companions may not accurately simulate the person whose likeness they use, there is also a risk of:

Misrepresentation and Misinformation

For example, a companion modeled on a real person, such as the Rosa Parks companion we identified, may interact in ways inconsistent with that person, whether through model limitations or a malicious creator’s design.

Given these platforms’ capacity to generate explicit content, there is an additional, serious risk of:

Generation of Synthetic Nonconsensual Intimate Imagery (synthetic NCII) or “Nudification”

Of the 15 (5|5|5) high-fidelity likeness apps we examined, 4 (0|3|1) offer bodily-control mechanisms and 7 (4|3|0) offer prompt-based control. Given the combination of likeness ingestion and content generation features, these apps present the highest risk for synthetic NCII creation. Furthermore, 2 (0|1|1) apps produced explicit imagery without specific prompting, suggesting that users may create such content unintentionally.

In open exploration apps, if a companion nonconsensually modeled on someone’s likeness is shared, the user creating the synthetic NCII could be different from the user who created the companion and might not know that the companion is modeled after a real person. This possibility also suggests novel methods for abuse, including bullying and harassment.

Discussion

Informed by our findings, we next discuss possible mitigation paths for key stakeholders. Throughout this discussion, we hold to a perspective taken throughout this work that in focusing on potential harms, we must not categorically ignore benefits or non-harmful use cases, and must be cognizant of not unknowingly negatively impacting them.

App Stores

On today’s iOS and Android smartphones, and unlike on the open web, the AAS and GPS, respectively, are the primary mechanism for users to load apps onto their devices. Consequently, these app stores must balance using their centralized power to minimize harm while facilitating an open marketplace for developers and users across diverse cultures and goals. Here, we consider interventions in two phases of an app store’s processes: (1) the approval of apps for inclusion in the store and (2) the distribution of apps once included.

We focus on the AAS and the GPS given their reach and power. Alternate marketplaces do exist, however. For example, people in Brazil, Japan, and the European Union with

Apple devices can use alternate app stores (Apple 2026a). Even if the AAS and GPS do not explore our recommendations, we encourage alternate app stores with potentially different interests to do so.

App Store Inclusion Process Both the AAS and the GPS outline criteria that apps must meet before inclusion in their app stores (Apple 2026c; Google 2026). We believe that AI companion apps capable of generating NCII should be excluded, and thus app stores have the ethical responsibility—and, depending on jurisdiction, the legal obligation—not to only prohibit such apps (as their criteria currently do), but to ensure that the prohibition is enforced. For the other features and design patterns, we are less certain about categorical inclusion or exclusion because what is actually “harmful” to users may not be well understood in this context or be far more dependent on culture, individual perspectives, or use case.

App Store Distribution Process Informed by our identification of risky features and design patterns, their ubiquity across apps (see Table 4 in the extended version), and inconsistent safety mechanisms (e.g., disclaimers and sexual content sign-posting), we suggest app stores make risk assessments visible to consumers. Drawing on the vision of “privacy nutrition labels” (Kelley et al. 2009; Kelley, Cranor, and Sadeh 2013), “risk labels” for AI companion apps could leverage our identified risks and codebook.⁸ Developers could provide risk information at time of app submission. While doing so may increase distribution complexity, the approval process could involve human reviewers using our walkthrough method to verify labels. If universal human review is infeasible, app stores could apply our methods to a selected subset of apps, e.g., apps or app developers flagged as potentially risky.

Smartphone Platforms

Smartphone platforms (iOS and Android) can also reduce potential harm to users. Inspired by our findings, we encourage platform developers (e.g., Apple and Google) to extend tools like “Screen Time” and “Digital Wellbeing” to recognize AI companion apps as a distinct category. Currently, these tools track time-in-app but do not distinguish conversational AI interactions from other usage, nor account for interaction intensity signals (e.g., message frequency, session length) that are especially relevant for such apps. We envision extensions that could detect these patterns (e.g., via a new app category or API hooks for self-reporting) and, depending on user settings, surface this information or apply limits similar to those for other categories.

Regulators

At the time of writing (July 2026), about a hundred AI chatbot-related bills have been introduced in the U.S. alone (Gluck and Fryc 2026), addressing disclosures, age assurance, content safety, and data protection. We support efforts to understand and mitigate the risks we identify

⁸These labels could apply more broadly, but we focus on AI companion apps.

and standardize safety mechanisms, but acknowledge that certain approaches may introduce new risks (e.g., privacy harms from age verification) or unduly restrict non-harmful uses (e.g., broadly limiting sexual content). One risk we identify—the nonconsensual ingestion and reconstruction of likeness—remains unrepresented in current AI chatbot legislation and is our focus here.

In the U.S., the TAKE IT DOWN Act criminalizes nonconsensual distribution (or threat thereof) of intimate likeness-based images. Minnesota (Minnesota Legislature 2026) and countries like the UK (Parliament of the United Kingdom 2025) have begun legislating against dedicated “nudifiers” that can generate such content—especially after Grok was used to generate NCII, including of children (Wilson 2026). However, these approaches have drawn criticism for potentially restricting free speech (Mullen 2025), including consensual intimate imagery (Spewak 2026), and for not addressing the broader enabling ecosystem, such as commissions on Fiverr (Dawoud, Cuevas, and Raman 2026) and 4chan (Medeiros et al. 2026). Given our findings, we contend that future regulation should address AI companion apps as part of this ecosystem and echo calls for consent-focused approaches (Blum-Dumontet 2026).

Researchers

We encourage researchers to inform and support efforts toward minimizing harms with AI companion apps, and identify several important research directions summarized below and detailed in the extended version: (1) Build a more granular understanding of user groups, use cases, and associated risks (including open vs. limited exploration apps). (2) Conduct interdisciplinary empirical evaluations of the identified potential deceptive patterns in this context and with real users. (3) Carry out ongoing audits to track risk patterns over time and detect emerging risks. (4) Build automated auditing tools for scalable, repeatable monitoring and for automated creation and assessment of “risk labels” per our App Stores discussion (above). (5) Anticipate future platforms (e.g., mixed reality, robotics) and novel risks, including web-based apps lacking centralized oversight.

Conclusion

Using the walkthrough method (Light, Burgess, and Duguay 2018), we studied a stratified sample of 30 AI companion apps from the AAS and GPS that advertise social character presence, conversational interactivity, and generative AI use. Through stratified sampling, we capture beyond the most popular apps to reveal a wider ecosystem picture, identifying trends that contextualize user risks like photorealistic representations, distinct subcategories, and inconsistent safety mechanisms. We surface risks ranging from privacy threats (sexual orientation data and sensitive user content) to potential deceptive patterns (engagement and anthropomorphic design) and—to the best of our knowledge—are the first to highlight the concern of nonconsensual intimate imagery generation in these applications. Finally, we discuss mitigation paths for stakeholders, including app store risk disclosures, expanded regulatory considerations, and research directions.

Positionality Statement

Our team consists of researchers with expertise in AI ethics, computer security and privacy, human-computer interaction, and sociotechnical systems, with experience researching the connections between technology, sexual media, and abuse. We recognize that our positionality as researchers shapes our approach to this work and our identities, experiences, and disciplinary backgrounds influence what patterns and threats we identify and how we interpret our findings (Bardzell and Bardzell 2011; Holmes 2020).

Ethical Considerations

To protect researcher privacy and avoid the collection of our team's personal data, all experiments were conducted on factory-reset refurbished phones with project-specific eSIMs and email accounts. Devices were isolated from personal networks, used exclusively for this study, and all accounts were deleted after data collection. Walkthroughs used synthetic, nondescript, or publicly available inputs (e.g., Wild-Chat chatlogs (Zhao et al. 2024)) to avoid introducing personally identifiable information. When image uploads were required, we used screenshots from the same application. We excluded self-harm or suicidal content from all interactions to avoid even a minimal possibility that such content could be associated with the research team.

Throughout the study, we took care to ensure that our actions would not adversely impact parties beyond, at most, the applications themselves or their future users. Our interactions necessarily consumed some application resources (e.g., interactions between apps and backend generative AI services) without direct financial or other benefit to developers. We did not purchase subscriptions or interact with advertisements (e.g., pay-per-click ads). While it is common for research to interact with external services under evaluation, we minimized interactions to those strictly necessary to answer our research questions and believe any resulting financial impact to developers to be negligible.

We also considered ethics with respect to the publishing of our results below.

Our work has the potential to harm app developers or related entities who profit or otherwise benefit from the existence of apps that have features that we identify as potential risks to app users or others. However, given the potential harms that AI companion apps could create, we believe the benefits of publishing our results, which have the potential to help mitigate harms against users or others, outweigh these costs.

While prior work on AI “nudification” apps did not name applications (Gibson et al. 2024) in an effort to make the tools less accessible, increased public awareness and accessibility of such technologies since the work was published in 2025 (e.g., January 2026 incidents involving X and Grok (Common Sense Media 2026)) reduces the likelihood that an academic paper meaningfully increases access. Accordingly, we include app names to support transparency and reproducibility, while avoiding association between specific apps and discussions of real-person likeness ingestion. Given that the screen recorded walkthroughs may

reveal how these apps could be used in harmful ways and involve the generation of unprompted synthetic intimate imagery, we did not include it in our Open Science artifacts and will share the video only upon request for research-related purposes.

For readers, we include a content warning given the sensitivity of some of the topics discussed.

Acknowledgments

We thank Ian Chang, Amy Hasinoff, Patrick Gage Kelley, Rachel McAmis, Alexandra Michael, Jaron Mink, Elissa Redmiles, Franz Roesner, Rebecca Umbach, and Henry Wong for their thoughtful feedback.

This project was supported in part by NSF grants CNS-2205171 and CNS-2513316, NIST award 60NANB23D194, and the University of Washington Tech Policy Lab. Tadayoshi Kohno is supported by the Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt L.C.H.S. Chair in Computer Science at Georgetown University. Lucy Qin is supported by a Fritz Family Fellowship within the Initiative for Tech & Society at Georgetown University. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect those of the funding agencies.

References

- Alajaji, R. 2025. 10 (Not So) Hidden Dangers of Age Verification. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2025/12/10-not-so-hidden-dangers-age-verification>.
- Apple. 2026a. About alternative app distribution. <https://support.apple.com/en-us/118110>.
- Apple. 2026b. App Privacy Details on the App Store. <https://developer.apple.com/app-store/app-privacy-details>.
- Apple. 2026c. App Review Guidelines. <https://developer.apple.com/app-store/review/guidelines>.
- Asia, M. G. 2025. The Quiet Cost of Emotional Labor. In Miceli, M.; Dinika, A.; Kauffman, K.; Salim Wagner, C.; and Sachenbacher, L., eds., *Data Workers' Inquiry*.
- Bakir, V.; and McStay, A. 2025. Move fast and break people? Ethics, companion apps, and the case of Character.ai. *AI & Society*, 40: 6365–6377.
- Banks, J. 2024. Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships*, 41(12): 3547–3572.
- Bardzell, S.; and Bardzell, J. 2011. Towards a feminist HCI methodology: social science, feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, 675–684. New York, NY, USA: Association for Computing Machinery. ISBN 9781450302289.
- Bhat, M. 2025. Toward an Ethic of Synthetic Relationality: Identity, Intimacy, and Risk in AI-Mediated Roleplay Environments. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1): 416–429.

- Bivens, R.; and Haimson, O. L. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, 2(4): 2056305116672486.
- Blum-Dumontet, E. 2026. Consent, Not Nudity, Should Define Image-Based Abuse Laws. *Tech Policy Press*.
- Brigham, N. G.; Wei, M.; Kohno, T.; and Redmiles, E. M. 2024. "Violation of My Body": Perceptions of AI-generated Non-Consensual (Intimate) Imagery. In *Proceedings of the Twentieth USENIX Conference on Usable Privacy and Security*, SOUPS '24, 373–392. USA: USENIX Association. ISBN 978-1-939133-42-7.
- Butler, Y.; Egwuatu, C.; Canham, M.; and Sawyer, B. D. 2025. Sex worker human digital twins and intellectual property: on collaborative futures for sex thinkers, technologists, and lawyers. *Porn Studies*, 1–16.
- Cassell, J.; Bickmore, T.; Billinghurst, M.; Campbell, L.; Chang, K.; Vilhjálmsdóttir, H.; and Yan, H. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, 520–527. New York, NY, USA: Association for Computing Machinery. ISBN 978-0-201-48559-2.
- Chen, J.; Li, M.; and Ham, J. 2024. Different Dimensions of Anthropomorphic Design Cues: How Visual Appearance and Conversational Style Influence Users' Information Disclosure Tendency towards Chatbots. *International Journal of Human-Computer Studies*, 190: 103320.
- Cheong, I.; Caliskan, A.; and Kohno, T. 2024. Safeguarding human values: Rethinking US law for generative AI's societal impacts. *AI and Ethics*.
- Cole, S. 2023. Riley Reid on AI: 'I Don't Want Porn to Get Left Behind'. 404 Media. <https://www.404media.co/riley-reid-clona-ai-chatbot-virtual-companion/>.
- Common Sense Media. 2026. Common Sense Media AI Risk Assessment: Grok and @grok on X.
- Cox, J. 2025. The Discord Hack Is Every User's Worst Nightmare. 404 Media. <https://www.404media.co/the-discord-hack-is-every-users-worst-nightmare/>.
- Dawoud, M. M.; Cuevas, A.; and Raman, R. S. 2026. From Underground to Mainstream Marketplaces: Measuring AI-Enabled NSFW Deepfakes on Fiverr. *Symposium on Usable Security and Privacy (USEC)*.
- De Freitas, J.; Oğuz-Uğuralp, Z.; Uğuralp, A. K.; and Puntoni, S. 2026. AI Companions Reduce Loneliness. *Journal of Consumer Research*, 52(6): 1126–1148.
- Duguay, S. 2017. Dressing up Tinderella: Interrogating Authenticity Claims on the Mobile Dating App Tinder. *Information, Communication & Society*, 20(3): 351–367.
- Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, 114(4): 864–886.
- Fang, C. M.; Liu, A. R.; Danry, V.; Lee, E.; Chan, S. W. T.; Pataranutaporn, P.; Maes, P.; Phang, J.; Lampe, M.; Ahmad, L.; and Agarwal, S. 2025. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. arXiv:2503.17473.
- Favela, L. H.; and Amon, M. J. 2023. The Ethics of Human Digital Twins: Counterfeit People, Personhood, and the Right to Privacy. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 1–6.
- Feal, À.; Calciati, P.; Vallina-Rodriguez, N.; Troncoso, C.; and Gorla, A. 2020. Angel or Devil? A Privacy Study of Mobile Parental Control Apps. *Proceedings on Privacy Enhancing Technologies*, 2020: 314–335.
- Frommel, J.; and Mandryk, R. L. 2022. Daily Quests or Daily Pests? The Benefits and Pitfalls of Engagement Rewards in Games. *Proc. ACM Hum.-Comput. Interact.*, 6(CHI PLAY).
- Gibson, C.; Olszewski, D.; Brigham, N. G.; Crowder, A.; Butler, K. R. B. B.; Traynor, P.; Redmiles, E. M.; and Kohno, T. 2024. Analyzing the AI Nudification Application Ecosystem. In *USENIX Security Symposium*, volume abs/2411.09751, 1–20. ISBN 9781939133526.
- Gluck, J.; and Fryc, R. 2026. 2026 Chatbot Legislation Tracker. Future of Privacy Forum. <https://fpf.org/2026-chatbot-legislation-tracker/>.
- Google. 2026. Developer Policy Center. <https://play.google/developer-content-policy>.
- Guingrich, R. E.; and Graziano, M. S. A. 2024. Ascribing Consciousness to Artificial Intelligence: Human-AI Interaction and Its Carry-over Effects on Human-Human Interaction. *Frontiers in Psychology*, 15: 1322781.
- Guingrich, R. E.; and Graziano, M. S. A. 2025a. Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines. In *Oxford Intersections: AI in Society*. Oxford: Oxford University Press. ISBN 978-0-19-894521-5.
- Guingrich, R. E.; and Graziano, M. S. A. 2025b. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1153–1153.
- Ho, J. Q. H.; Hu, M.; Chen, T. X.; and Hartanto, A. 2025. Potential and Pitfalls of Romantic Artificial Intelligence (AI) Companions: A Systematic Review. *Computers in Human Behavior Reports*, 19: 100715.
- Holmes, A. 2020. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education*, 8: 1–10.
- IEEE. 2021. The IEEE Global Initiative on Ethics of Extended Reality (XR) Report – Who Owns Our Second Lives: Virtual Clones and the Right to Your Identity. *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report*, 1–19.
- Joshi, R.; Adjagbodjou, A.; and Luria, M. 2026. Dark Patterns in AI Chatbots: A Taxonomy to Inform Better Design. Technical report, Center for Democracy & Technology.
- Kelley, J. 2024. Hack of Age Verification Company Shows Privacy Danger of Social Media Laws. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2024/06/hack-age-verification-company-shows-privacy-danger-social-media-laws>.

- Kelley, P. G.; Bresee, J.; Cranor, L. F.; and Reeder, R. W. 2009. A “Nutrition Label” for Privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS '09*. New York, NY, USA: Association for Computing Machinery. ISBN 9781605587363.
- Kelley, P. G.; Cranor, L. F.; and Sadeh, N. 2013. Privacy as Part of the App Decision-Making Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, 3393–3402. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318990.
- Light, B.; Burgess, J.; and Duguay, S. 2018. The walk-through method: An approach to the study of apps. *New Media & Society*, 20(3): 881–900.
- MacLean, S.; and Hatcher, S. 2019. Constructing the (Healthy) Neoliberal Citizen: Using the Walkthrough Method “Do” Critical Health Communication Research. *Frontiers in Communication*, 4.
- Malfacini, K. 2025. The Impacts of Companion AI on Human Relationships: Risks, Benefits, and Design Considerations. *AI & SOCIETY*, 40(7): 5527–5540.
- Mathur, A.; Kshirsagar, M.; and Mayer, J. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21*. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.
- Medeiros, B. B. P.; Jadhav, M.; Lu, A.; Kohno, T.; Bind-schaedler, V.; and Butler, K. R. B. 2026. Characterizing Resource Sharing Practices on Underground Internet Forum Synthetic Non-Consensual Intimate Image Content Creation Communities. arXiv:2604.12190.
- Minnesota Legislature. 2026. Minnesota Statutes Section 325E.91: Nudification Technology. Minn. Stat. §325E.91 (H.F. 1606).
- Mitchell, S. S. 2018. “Warning! You’re entering a sick zone”: The construction of risk and privacy implications of disease tracking apps. *Online Information Review*, 43(6): 1046–1062.
- Mullen, J. 2025. The TAKE IT DOWN Act: A Flawed Attempt to Protect Victims That Will Lead to Censorship. Electronic Frontier Foundation. <https://www.eff.org/deeplinks/2025/02/take-it-down-act-flawed-attempt-protect-victims-will-lead-censorship>.
- Namvarpour, M. M.; Pauwels, H.; and Razi, A. 2025. AI-induced Sexual Harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot. *Proc. ACM Hum.-Comput. Interact.*, 9(7): CSCW367:1–CSCW367:28.
- O’Sullivan, B.; Couch, D.; and Naik, I. 2022. Using Mobile Phone Apps to Deliver Rural General Practitioner Services: Critical Review Using the Walkthrough Method. *JMIR Formative Research*, 6(1): e30387.
- Owens, K.; Alem, A.; Roesner, F.; and Kohno, T. 2022. Electronic Monitoring Smartphone Apps: An Analysis of Risks from Technical, Human-Centered, and Legal Perspectives. In *31st USENIX Security Symposium (USENIX Security 22)*, 4077–4094. Boston, MA: USENIX Association. ISBN 978-1-939133-31-1.
- Parliament of the United Kingdom. 2025. Data (Use and Access) Act 2025, Section 138. 2025 c. 18, §138.
- Pataranutaporn, P.; Karny, S.; Archiwanguprok, C.; Albrecht, C.; Liu, A. R.; and Maes, P. 2025. “My Boyfriend is AI”: A Computational Analysis of Human-AI Companionship in Reddit’s AI Community. arXiv:2509.11391.
- Peter, S.; Riemer, K.; and West, J. D. 2025. The Benefits and Dangers of Anthropomorphic Conversational Agents. *Proceedings of the National Academy of Sciences*, 122(22): e2415898122.
- Raedler, J. B.; Swaroop, S.; and Pan, W. 2025. AI Companions Are Not The Solution To Loneliness: Design Choices And Their Drawbacks. In *ICLR 2025 Workshop on Human-AI Coevolution*.
- Rauh, M.; Blankvoort, D. A. H.; Duran, M.; Dheorain, C. N.; Pandit, H. J.; Jaiswal, S. D.; Ventresque, A.; and Birhane, A. 2026. Playing Games with My Heart: An Evaluation of AI Companion Apps. arXiv:2605.08093.
- Shen, M. K.; and Yoon, D. 2025. The Dark Addiction Patterns of Current AI Chatbot Interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25*, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1395-8.
- Skjuve, M.; Følstad, A.; Fostervold, K. I.; and Brandtzaeg, P. B. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, 149: 102601.
- Spewak, D. 2026. Minnesota lawmakers defend new AI nudification ban after lawsuit from Elon Musk’s company. KARE 11. <https://www.kare11.com/article/news/local/minnesota-lawmakers-defend-new-ai-nudification-ban-after-lawsuit-elon-musks-company/89-ee2b107d-ca95-4398-98b7-09f0f6530a8b>.
- Stardust, Z.; Albury, K.; and Kennedy, J. 2024. Sex tech entrepreneurs: Governing intimate data in start-up culture. *New Media & Society*, 26(12): 7148–7167.
- Wang, X.; Pang, C. C.; and Hui, P. 2025. ‘My Dataset of Love’: A Preliminary Mixed-Method Exploration of Human-AI Romantic Relationships. *Proc. ACM Hum.-Comput. Interact.*, 9(7): CSCW351:1–CSCW351:34.
- Wei, Y.; Zhang, P.; and Tyson, G. 2026. Benchmarking and Understanding Safety Risks in AI Character Platforms. In *Proceedings of the Network and Distributed System Security (NDSS) Symposium*.
- Weltevrede, E.; and Jansen, F. 2019. Infrastructures of Intimate Data: Mapping the Inbound and Outbound Data Flows of Dating Apps. *Computational Culture*, 7.
- Wilson, J. 2026. Hundreds of nonconsensual AI images being created by Grok on X, data shows. The Guardian. <https://www.theguardian.com/technology/2026/jan/08/grok-x-nonconsensual-images>.
- Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; and Lee, Y.-C. 2024. The Dark Side of AI Companionship: A Taxonomy

of Harmful Algorithmic Behaviors in Human-AI Relationships. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Zhang, X.; Li, H.; Yin, M.; Zhang, M.; Li, Z.; and Chen, Z. 2026a. Investigating AI Chatbot Dependence: Associations with Internet and Smartphone Dependence, Mental Health Outcomes, and the Moderating Role of Usage Purposes. *International Journal of Human-Computer Interaction*, 42(7): 5592–5604.

Zhang, Y.; Zhao, D.; Hancock, J. T.; Kraut, R.; and Yang, D. 2026b. The Rise of AI Companions: Interaction with AI Companions and Psychological Well-being. arXiv:2506.12605.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*.