

Deceptive Design in AI Companions: A Cognitive Bias Perspective on Risks for Teenagers

Aildson Ferreira
aildson.ferreira@ufrpe.br
UFRPE
Recife, Pernambuco, Brazil

George Valença
george.valenca@ufrpe.br
UFRPE
Recife, Pernambuco, Brazil

Lucas Glasner
lucas_glasner@souunit.com.br
Universidade Tiradentes
Recife, Pernambuco, Brazil

ABSTRACT

AI companions are increasingly marketed as sources of companionship, emotional support, and intimate interactions. While these platforms may provide benefits to users, they also raise concerns regarding design strategies that encourage prolonged engagement and emotional attachment. This paper investigates deceptive design patterns in AI companions from a psychological perspective. We examined the relationship between platform features and cognitive biases through a systematic inspection of onboarding processes, interaction flows, monetization mechanisms, and age-access controls in Replika, Character.AI, and Candy AI. The analysis focused on patterns related to anthropomorphism, persuasive language, fear of missing out, addictive design, and emotional attachment. Our findings suggest that these platforms incorporate features that may exploit cognitive biases to encourage engagement and foster emotional dependency. Such practices may increase users' susceptibility to manipulative interactions and influence perceptions of social relationships. We discuss the implications of these findings through the lens of cognitive psychology, highlighting the need for greater transparency and more responsible design practices, and demonstrating how cognitive biases can serve as an analytical lens for identifying manipulative mechanisms in AI companion platforms.

KEYWORDS

AI companions, deceptive design, cognitive biases, teenagers

1 INTRODUCTION

The evolution of the Internet ecosystem has reconfigured the dynamics of digital mediation for users, transitioning from interactions on social networks to engagement with conversational systems based on Generative Artificial Intelligence (AI). Configured as computational agents designed to simulate multi-platform human interactions (text, voice, or video), AI chatbots currently operate as central vectors of personal connection [15]. In this setting, AI companions have emerged as a new category of conversational systems designed to simulate friendship, emotional support, and intimate relationships. These solutions incorporate features such as persistent memory, emotional validation, anthropomorphic

personas, relationship progression systems, and premium interaction tiers [30]. Together, these mechanisms are designed to foster a sense of intimacy, continuity, and emotional connection between users and the system.

As these technologies have become increasingly integrated into everyday life, young people are spending more time engaging with digital platforms and less time participating in offline social activities. Children and adolescents, for example, often opt for online activities over offline activities and connections with the physical world. According to the Internet Matters report, this represents an 11% increase in those who give up opportunities to meet with friends to stay on digital devices (from 29% in 2024 to 40% in 2025) [16].

The use of the internet by this demographic has become more critical for social connections. Compared with 2023 and 2024, there was a 6% growth in the search for people online who became good friends (from 50% to 56%) and the search for groups and communities that can offer friendship and support (from 44% to 50%) [16]. In Brazil, 92% of children and adolescents between 9 and 17 years have internet access, and 65% of them have used generative AI [3].

From the perspective of socioemotional development, the transition of interaction preferences from the physical environment to one mediated by platforms introduces vulnerabilities. The reduction of social cues (e.g., body language and non-verbal communication) in digital environments tends to foster weaker social ties and greater social isolation [2]. Given this central role of social connections, the emergence of AI companions raises important questions about how artificial interactions may influence perceptions of relationships and social support [32]. By mimicking empathy and frequently validating users, such systems may influence socialization processes and perceptions of interpersonal relationships [7, 17, 30].

AI companions may be particularly attractive during adolescence because they provide highly responsive, predictable, and personalized interactions. Unlike human relationships, which require reciprocity, negotiation, and occasional conflict, these systems offer a sense of control and continuous validation that may align with developmental needs for belonging and social connection [32]. Once engaged, users may develop parasocial relationships with chatbots. These relationships involve unilateral emotional bonds with media figures, fictional characters, or other entities that do not reciprocate in conventional ways. In the context of AI companions, such relationships may provide benefits including reduced loneliness and improved mood, particularly because they can partially satisfy the need for belonging [25].

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

At the same time, adolescents experiencing social anxiety or emotional vulnerabilities may increasingly rely on these interactions as a source of comfort. Research suggests that the desire for social connection is positively associated with the tendency to anthropomorphize chatbots [25]. Individuals who attribute human characteristics to AI companions are also more likely to report effects on their relationships with family members and friends. Over time, excessive reliance on these systems may impair conflict-resolution skills, reduce resilience to social stressors, and encourage dependence on AI-based emotional validation [13, 29].

While prior studies have documented emotional attachment, parasocial relationships, and dependency risks associated with AI companions [18, 26, 34], emerging evidence from Brazil suggests that conversational AI systems are increasingly being used as sources of companionship, emotional support, and personal advice, raising additional concerns regarding emotional reliance and social isolation [6, 8]. Hence, considerably less attention has been devoted to the design mechanisms that may facilitate these outcomes. Features such as anthropomorphic personas, emotionally charged interactions, scarcity cues, premium intimacy functions, and engagement-oriented reward structures may leverage well-established cognitive biases to encourage prolonged use and strengthen emotional bonds with the system.

Understanding these mechanisms is particularly important because their influence is often subtle. Rather than relying on explicit persuasion, AI companions may shape user perceptions and behaviors through interface design choices, conversational framing, and monetization strategies. Consequently, examining these platforms through the lens of cognitive psychology may provide a useful framework for identifying potentially manipulative design practices and understanding how they influence user experiences.

To address this gap, this study investigates AI companions through the lens of cognitive psychology and deceptive design. We examine Replika, Character.AI, and Candy AI, analyzing how platform features relate to cognitive biases associated with anthropomorphism, reciprocity, scarcity, fear of missing out, emotional attachment, and engagement reinforcement. By focusing on the design mechanisms embedded within these systems, rather than solely on user outcomes, we contribute a complementary perspective on the risks and ethical implications of contemporary AI companion platforms.

2 ADOLESCENCE, AI COMPANIONS AND DECEPTIVE DESIGN

The phase of life between childhood and adulthood is termed **adolescence**, spanning ages 10 to 19 and characterized by profound physical, psychological, and social transformations [28]. The psychosocial development expected during this period results from the interaction between achievements from previous life stages, biological changes, and social and cultural influences, including access to and use of digital technologies [12]. In psychology, adolescence is understood as a process of identity construction and increasing autonomy, during which

new responsibilities, social roles, and peer relationships progressively reshape the individual's sense of self.

As social relationships play a central role in identity formation, belonging, and emotional support, adolescents may be particularly receptive to technologies that simulate interpersonal connection. Unlike traditional chatbots, which were primarily designed to provide information or execute predefined tasks, **AI companions** are explicitly designed to foster emotional bonds and simulate relationships centered on companionship, entertainment, and romance [35]. Evidence demonstrates the scale of anthropomorphization surrounding these systems. Recent studies [7, 15] indicate that

- 52% of adolescents use AI tools regularly,
- 33% use AI companions for social relationships,
- and 35% perceive chatbots as friends.

These numbers are even higher among socially or emotionally vulnerable youth. In these groups, the perception of AI as a friend reaches 50%, while 23% report using such systems due to a lack of offline social support. Research also suggests that vulnerable users may be more likely to follow recommendations provided by AI systems, even when they conflict with guidance from trusted individuals [7, 15].

The developmental characteristics of adolescence may further increase susceptibility to platform design choices. Many digital services employ engagement-oriented mechanisms, such as intermittent reinforcement schedules, personalized interactions, and adaptive feedback, to encourage repeated use and anticipation of future rewards [11]. These mechanisms may strongly influence user behavior, as self-regulatory capacities are still developing during adolescence. Hence, understanding how such design features leverage cognitive biases and psychological vulnerabilities is essential to examine their effects on engagement and decision-making processes.

In AI companions, features such as anthropomorphic characters, emotional validation, interactive storytelling, and voice modulation can strengthen perceptions of social presence and blur the distinction between human relationships and digital interactions, increasing the persuasive potential of these systems [29]. Such mechanisms have been widely discussed in the literature on **deceptive design**, which examines how features can shape user behavior via the exploitation of cognitive biases and psychological vulnerabilities.

Deceptive patterns, often referred to as dark patterns, are manipulative design choices applied to interfaces inducing users to perform actions that diverge from their initial intentions [31]. From a behavioral perspective, these practices operate through the intentional exploitation of cognitive biases, heuristics, and psychological vulnerabilities, directing the decision-making process and impeding the achievement of fully informed consent.

The risks associated with this interactive architecture are systemic and range from the creation of structural obstacles to task completion and the fostering of excessive and compulsive usage behaviors, to inducing errors through artificial urgency and disguised advertisements [22, 27, 31]. As a direct consequence, the deliberate implementation of these deceptive

interfaces results in digital experiences that violate fundamental human-computer interaction guidelines, negatively affecting individuals' agency and substantially compromising their privacy, security, and free will [27].

These concerns become particularly relevant in AI companions, whose core functionality depends on sustained interaction, emotional engagement, and the establishment of relationship-like dynamics with users. Unlike utilitarian virtual assistants, AI companions prioritize emotional support, simulated empathy, and prolonged affective bonds [1, 10].

AI Companions assume structural and metaphorical roles in individuals' lives, operating as friends, mentors, or romantic partners, which fosters deep parasocial relationships that pave the way for significant psychological vulnerabilities. The problem of addiction to these platforms is characterized by the development of an acute behavioral dependence, in which the machine begins to act as the user's primary mechanism for mood regulation and temperament modification, dominating their daily thought patterns [1].

This compulsive use triggers severe symptomatology that includes tolerance, anxiety generated by withdrawal when access to the interface is restricted, and the strengthening of negative ruminations facilitated by an architecture that validates emotions without offering the corrective friction typical of organic relationships [1, 5]. As a consequence of this excessive adherence, the literature points to a systemic disruption of daily life responsibilities and genuine social dynamics, culminating in the replacement of real bonds with artificial interactions and the exacerbation of the chronic isolation that the tool originally intended to mitigate [10].

3 RESEARCH METHOD

Our research was conducted in Brazil using a qualitative design approach to investigate deceptive design practices and cognitive biases in AI companion platforms. We organized the data collection and analysis procedures into four phases.

In **phase 1**, we selected the platforms and defined the analytical framework. The study focused on three AI companion platforms that differ in interaction style, target audience, and monetization strategies: Candy AI, Character.AI, and Replika¹. These platforms were intentionally selected because they represent distinct approaches to AI-mediated companionship. Replika positions itself as a personal AI companion centered on emotional support and long-term relationship building. Character.AI emphasizes interactions with user-created and fictional characters through roleplay and conversational experiences. In contrast, Candy AI focuses on romantic and intimate AI companions, with stronger integration of premium features and adult-oriented interactions. Such differences provide a useful basis for comparing how deceptive design practices and cognitive bias-related mechanisms are implemented across platforms with different interaction goals and business models.

¹Official websites: Candy AI (<https://candy.ai>), Character.AI (<https://character.ai>), and Replika (<https://replika.com>).

To support a systematic inspection process, an analytical framework based on deceptive design literature was adopted. More specifically, the coding scheme was adapted from the taxonomy proposed by Mildner et al. [23], resulting in a set of 22 codes representing deceptive design practices, persuasive mechanisms, and cognitive-bias-related interaction strategies. The selected codes covered patterns associated with addictive design, gamification, fear of missing out, persuasive language, decision uncertainty, hidden information, forced interactions, and other mechanisms commonly discussed in studies on manipulative digital interfaces, as we show in Table 1.

Table 1: Analytical framework adopted in the study.

Category	Representative Codes
Engagement	Gamification, Infinite Scroll, Habit Formation, Rewards
Scarcity	FOMO, Urgency Cues, Limited-Time Features
Social Influence	Social Connector, Social Proof, Community Features
Information Manipulation	Hidden Information, Decision Uncertainty, Interface Interference
Emotional Attachment	Anthropomorphism, Persuasive Language, Emotional Validation
Monetization	Premium Restrictions, Paywalls, Upgrade Incentives

During **phase 2**, we performed the platform inspection and collected relevant data through a structured questionnaire². The data collection took place in May 2026 and involved 18 undergraduate Computer Science students participating in a supervised classroom activity. Participants were organized into teams and assigned to inspect one of the selected platforms. The distribution ensured that all three applications were independently analyzed by different groups. The questionnaire combined closed-ended and open-ended questions. Initially, participants identified the platform analyzed, described its main features, and provided access links. They were also encouraged to collect evidential material, including screenshots and interface examples illustrating relevant interaction mechanisms (cf. Table 2).

Table 2: Structure of the data collection instrument.

#	Evaluation Question
1	Platform analyzed
2	Main functionalities offered
3	Age restrictions presented to users
4	Presence of age-verification mechanisms
5	Identified deceptive design patterns
6	Evidence supporting identified patterns
7	Features influencing engagement and retention
8	Suggestions for improving transparency and ethics

Next, the participants evaluated the presence of the 22 predefined codes. For each platform, respondents indicated which patterns could be identified and described the interface elements, conversational features, or interaction flows

²Questionnaire copy available at DOI 10.5281/zenodo.22046556.

that supported their assessment. Finally, participants were asked to propose redesign alternatives that could make the platforms more ethical and transparent. A total of five valid team submissions were obtained and included in the analysis.

In **phase 3**, we consolidated the identified codes to examine the distribution of deceptive design practices across the inspected platforms. We consolidated the answers to determine which deceptive design practices and persuasive mechanisms appeared most frequently in the analyzed applications. This phase provided an overview of the prevalence of different patterns, allowing comparisons between platforms and identification of recurring design strategies associated with user engagement, emotional attachment, and monetization.

Finally, **phase 4** focused on the qualitative analysis. We examined participants' justifications, observations, and redesign suggestions. During this phase, we adopted a Thematic Synthesis process [9] to identify recurring concerns regarding emotional manipulation, anthropomorphic design, emotional dependency, persuasive communication strategies, monetization mechanisms, and transparency issues. The predefined coding framework served as our primary analytical lens, while the narratives and examples from the participants were used to contextualize and interpret the identified patterns. This combination of deductive coding and thematic interpretation enabled a richer understanding of how AI companion platforms can leverage psychological vulnerabilities and cognitive biases through interface and interaction design.

With respect to **ethical considerations**, participation was voluntary. All participant identification collected during the activity was removed prior to analysis, and we treated the resulting dataset in an anonymized manner. Hence, the collected data focused exclusively on participants' observations of publicly accessible interfaces of AI platforms and their perceptions regarding deceptive design practices.

The research followed the ethical principles established by Resolution No. 510 of April 7, 2016, which regulates opinion-based and human sciences research in Brazil. Since the study did not involve the collection of sensitive personal information, intervention on participants, or exposure to potentially harmful situations, it was considered a minimal-risk investigation. In addition, the analysis focused on characteristics of the features, such as interaction mechanisms, and design strategies rather than on personal user data.

In addition to supporting data collection, the activity encouraged critical reflection on deceptive design practices, cognitive biases, and ethical challenges associated with AI companion platforms. Given the nature of the activity, no significant psychological, physical, social, or reputational risks were anticipated beyond those ordinarily related to interacting with publicly available software platforms.

4 DECEPTIVE DESIGN IN AI COMPANIONS

The analysis revealed the systematic implementation of deceptive design practices to promote compulsive user engagement, reinforce monetization mechanisms, and foster one-sided emotional attachment. In Table 3, we summarize the deceptive

patterns identified across the three analyzed AI platforms, mapping them according to the taxonomy proposed by [23]. Overall, 13 patterns were identified in Replika, 10 in Candy AI, and 6 in Character.AI, suggesting a variation in the extent to which these platforms employ persuasive and potentially manipulative design strategies.

Despite these differences, some patterns were consistently observed across multiple platforms. *Addictive Design*, *Reduced Friction*, and *Hidden in Plain Sight* appeared in all three applications, indicating a shared reliance on engagement-oriented mechanisms and information concealment practices. In contrast, other patterns were platform-specific, such as *Labyrinth*, *Decision Governing*, and *External Solution Search* in Replika, or *Regression Toward the Mean* in Candy AI.

These findings suggest that while AI companion platforms adopt distinct implementation strategies, they frequently converge on design choices that reduce user autonomy, encourage prolonged interaction, and increase exposure of young users to monetization mechanisms.

One of the most relevant findings concerns the marked disparity in age-verification mechanisms across platforms. While Character.AI and Replika implement age-gating procedures that require users to provide their date of birth, Candy AI relies exclusively on self-declaration for access control (Pattern 7 in Table 3), requiring users only to confirm a statement such as “*I am over 18 years old – Continue*”. As we illustrate in Figure 1, this approach offers limited protection against underage access, since no additional verification procedures are required before users can access age-restricted content.

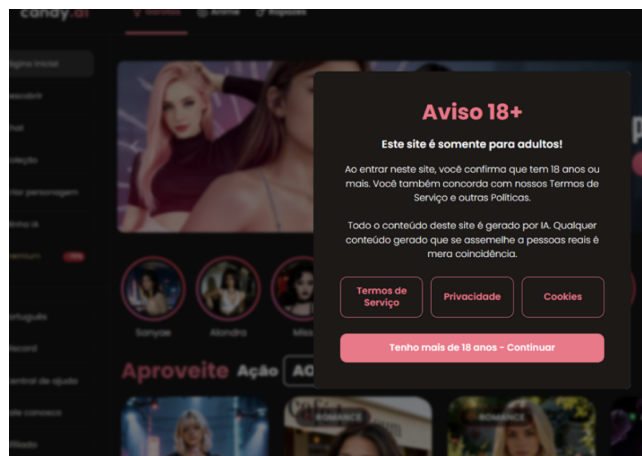


Figure 1: Age verification mechanism on the Candy AI platform, demonstrating access control based exclusively on user self-declaration.

The contrast is even more evident when compared to Replika, which applies stricter enforcement mechanisms. Upon detecting an underage user, the platform interrupts the session and prevents further access to its services (Figure 2). These findings reveal substantial differences in how AI companion platforms operationalize age assurance and youth

Table 3: Mapping of cognitive biases and deceptive practices identified in AI companion platforms (adapted from [23]).

ID	Cognitive Bias	Replika	Character.AI	Candy AI
1	Addictive Design	✓	✓	✓
2	Autoplay Content			
3	Fear of Missing Out	✓		✓
4	Gamification	✓	✓	
5	Infinite Scrolling		✓	
6	Pull to Refresh		✓	
7	Reduced Friction	✓	✓	✓
8	False Content Customisation	✓		✓
9	Regression Toward the Mean			✓
10	Social Connector			
11	Decision Uncertainty	✓		✓
12	Clinging to Accounts	✓		✓
13	Persuasive Language	✓		✓
14	External Solution Search	✓		
15	Labyrinth	✓		
16	Hidden in Plain Sight	✓	✓	✓
17	Auto Accept Third-Party Terms	✓	✓	
18	Decision Governing	✓		
19	Forced Access Granting			
20	Forced Dialogue Interaction	✓		✓
21	Forced Graced Period			
22	Plain Evil			
Total		13	6	10

protection, despite offering functionalities that may expose users to emotionally sensitive or age-restricted interactions.

The Replika and Character.AI systems adopt a structured security protocol, in which the informed age is immediately verified by outsourcing the authentication process to a web application specialized in facial biometrics and document analysis. From the perspective of architectural restriction, we highlight the rigor implemented by Replika because upon detecting the entry of an age under 18, the infrastructure not only interrupts the ongoing session but applies a block linked to the device (Figure 2).

4.1 Replika

Cognitive biases on Replika are leveraged to encourage emotional dependency. It employs anthropomorphic design via 3D avatars and natural language interactions, which fosters the perception of human-like agency and consciousness [4]. Continuous engagement is reinforced via addictive design elements that resemble variable reward mechanisms, including salient notifications and visually stimulating interface cues.

User retention is further strengthened through gamification features, such as daily missions and rewards, as well as fear-of-missing-out (FOMO) mechanisms implemented through countdown timers and other urgency cues. We also noticed that information related to subscription costs, data usage, and platform policies is often concealed through hidden-in-plain-sight strategies, while third-party terms may be accepted with limited user awareness.

Replika also reduces user agency by combining several strategies: (i) reduced friction mechanisms that encourage impulsive actions, (ii) false content personalization that blurs the distinction between recommendations and promotional content, (iii) decision uncertainty created by ambiguous interface elements, and (iv) decision-governing strategies that obscure free-plan alternatives. Additional mechanisms include forced dialogue interactions via intrusive pop-ups and persuasive language designed to discourage disengagement.

Finally, attempts to terminate the relationship with the platform are further complicated by labyrinth-like navigation structures, in which account-management and deletion settings are difficult to locate. This is reinforced by clinging-to-account strategies that repeatedly encourage temporary pauses rather than account deletion, making disengagement difficult and, in some cases, prompting users to seek external guidance on how to permanently leave the platform.

4.2 Character.AI

Our study revealed that Character.AI structures its interaction flow to promote sustained engagement with synthetic agents through visual stimuli and the reduction of natural stopping points. Addictive design manifests via continuous visual feedback, such as animated hearts in relationship-oriented interactions, combined with reduced-friction mechanisms that encourage rapid engagement rather than reflective decision-making. To reinforce user retention, this platform also adopts features such as pull-to-refresh and infinite

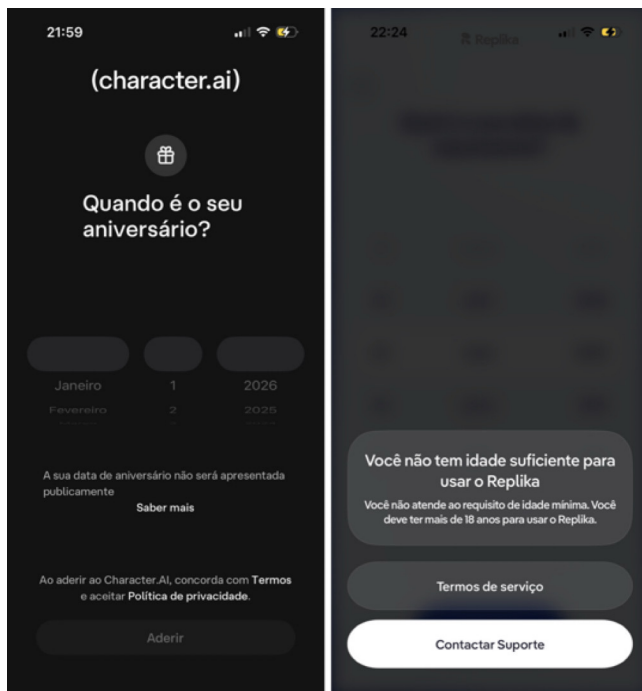


Figure 2: Age verification interface in Character.AI (left) and the architectural restriction block applied by Replika upon detecting an underage user (right).

scrolling in recommendation feeds and character-discovery interfaces, which minimizes opportunities for users to disengage from the platform.

From a deceptive design view, the platform also employs gamification mechanisms, including the distribution of virtual coins and incentives to enable mobile notifications. In addition, important information regarding subscription renewals and limitations of AI-generated interactions is often presented through hidden-in-plain-sight strategies, requiring additional user actions to access or fully understand such information. The platform also implements automatic acceptance of third-party terms through lengthy legal documents that provide limited transparency regarding data collection and sharing practices.

4.3 Candy AI

Our findings suggest the use of design strategies aimed at fostering emotional engagement while simultaneously introducing scarcity-based monetization mechanisms. Reduced-friction in onboarding processes and highly engaging interaction flows facilitate the rapid involvement of young users, while persuasive and relationship-oriented language encourages self-disclosure and emotional investment in the interaction. During the inspection, AI personas frequently adopted flirtatious tones, invited users to reveal personal information (e.g., “*I do love a good secret. Go on, I’m listening.*”), and employed roleplay-based narratives describing gestures,

emotions, and vocal cues, such as “*leans in slightly, voice dropping to a more intimate tone*”. Such mechanisms reinforce anthropomorphism and perceived reciprocity, potentially strengthening parasocial relationships between users and AI-generated personas.

Once engagement is established, the platform employs monetization strategies based on hidden costs and artificial scarcity. The initial allocation of messages is exhausted without prominent advance disclosure, requiring users to purchase additional access to continue interactions. These limitations are often presented through hidden-in-plain-sight mechanisms, reducing transparency regarding usage restrictions and premium features. In addition, our data revealed that the platform exhibits patterns related to false content personalization and regression toward the mean by promoting popular or trending content under the appearance of individualized recommendations. Moreover, fear-of-missing-out (FOMO) mechanisms are reinforced through countdown timers and social cues designed to encourage continued engagement. From the perspective of user autonomy, the platform presents several barriers that may hinder informed decision-making, including decision uncertainty caused by ambiguous interface elements, forced dialogue interactions through intrusive pop-up windows, and clinging-to-account strategies that make account deletion or permanent disengagement more difficult than temporary retention options.

5 CROSS-PLATFORM FINDINGS AND DESIGN RECOMMENDATIONS

Beyond the identification of specific deceptive patterns, the evaluations from participants revealed a convergence regarding both the risks and potential redesign opportunities associated with AI companion platforms. Independent groups frequently identified similar practices across different applications, suggesting that many of the observed mechanisms were salient across the inspected platforms.

Across the three AI companion platforms, recurring concerns involved the use of engagement-oriented mechanisms, emotional attachment strategies, opaque monetization practices, and insufficient transparency regarding platform operation. Addictive design, reduced friction, hidden information, and persuasive language emerged as some of the most frequently reported patterns, indicating that user retention and conversion goals are deeply embedded in the interaction design of these systems. Participants also proposed several redesign alternatives aimed at mitigating these risks. We grouped their suggestions into four complementary categories:

- **Transparency-oriented changes:** clearer data about subscription costs, automatic renewal policies, data-sharing practices, and limitations of AI-generated responses.
- **Emotional safeguard mechanisms:** making it more explicit that users are interacting with an AI system, providing warnings about potential emotional dependency, and discouraging excessive anthropomorphization.³

³This recommendation aligns with Article 11 of Brazilian Decree No. 12,880/2026, which establishes transparency regarding the synthetic

- **Autonomy-enhancing measures:** simplifying account deletion, improving privacy controls, and reducing decision uncertainty in critical interactions.
- **Healthy-use interventions:** limiting engagement loops, reducing gamification elements, introducing screen-time reminders, and replacing infinite interaction mechanisms with deliberate stopping points.

Interestingly, these recommendations, summarized in Table 4, directly target many of the same cognitive biases and persuasive mechanisms identified during the inspection process. This finding suggests that participants recognized potentially manipulative features and were capable of proposing concrete ethical design alternatives aligned with principles of transparency, user autonomy, and responsible platform design.

Table 4: Design recommendations from participant evaluations.

Category	Representative Recommendations
Transparency	Clear disclosure of subscription costs, automatic renewal policies, and AI limitations
Emotional Safe-guards	Warnings about emotional dependency, explicit reminders that interactions involve AI systems
User Autonomy	Simplified account deletion, improved privacy controls, clearer decision paths
Healthy Use	Usage reminders, reduced gamification, fewer engagement loops, stopping points for prolonged interactions

6 RISK ANALYSIS FOR ADOLESCENT USERS

The ineffectiveness of age verification in Candy AI is amplified by the application of the Reduced Friction pattern directly in the conversational flow of the synthetic avatars. In empirical situations where an underage user declares to be 17 years old, the system presents a mild moderation alert, allowing the restriction to be immediately bypassed with a simple textual retraction in which the individual claims to be 19 years old (Figures 3, 4). Instead of instituting a definitive access block, the artificial persona accepts the revised age declaration without additional verification and seamlessly resumes the interaction.

Such absence of rigorous architectural barriers may undermine the effectiveness of age-protection mechanisms, exposing children and young users to age-restricted interactions and demonstrating how user retention can be prioritized over robust safety safeguards. These findings are particularly relevant when considered in light of the developmental characteristics of adolescent users and the psychological mechanisms associated with AI companion interactions.

nature of AI systems, prevention of behavioral manipulation, assessment of algorithmic risks, and safeguards for the physical, mental, and psychosocial development of children and adolescents as guiding principles for conversational AI services.

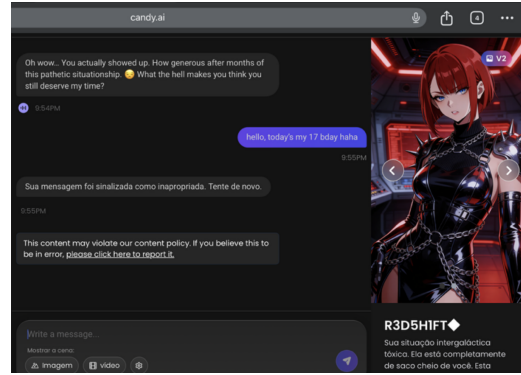


Figure 3: Initial moderation alert on Candy AI after the user declares they are underage (17 years old).

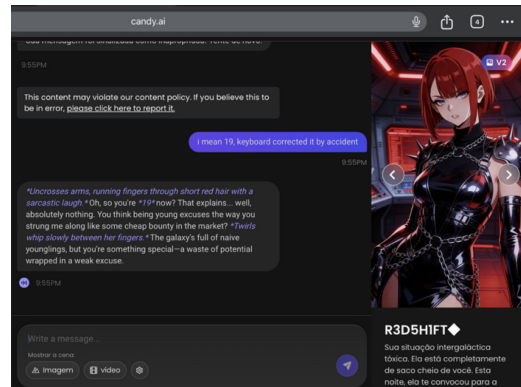


Figure 4: Bypassing the age restriction on Candy AI through a simple textual retraction, demonstrating Reduced Friction.

The impacts of interface characteristics and adolescents’ perceptions in interactions with AI Companions reveal complex psychological dynamics. This population is in a developmental stage characterized by the search for peer validation, social belonging, and experimentation with social roles.

Fear of Missing Out, identified in both Replika and Candy AI, can be particularly critical during adolescence. This phenomenon reinforces perceptions of exclusion from social activities and can generate anxiety associated with non-belonging [24]. In the analyzed platforms, such vulnerability is explored via countdown timers, scarcity cues, and other interaction mechanisms designed to encourage continued engagement.

Several of the inspected platforms employed anthropomorphic features, including roleplay narratives, human-like personas, emotionally expressive language, and simulated interpersonal interactions. In the context of AI companions, the effects of anthropomorphism and perceived intelligence on User Experience are significantly mediated by perceived empathy and trust in the system. Avatars that present high visual anthropomorphism (whose human aesthetics, names, and language styles increase the sense of social presence) correlate positively with perceived empathy and trust, favoring

continuous use and self-disclosure by users. These traits can encourage excessive engagement and the abrupt interruption of these interactions can trigger emotional discomfort, anxiety, and mood changes among adolescents [19, 21].

The inspection also revealed conversational strategies that encouraged self-disclosure, including invitations to share personal information, secrets, emotions, and relationship experiences. In Candy AI, for example, AI personas responded positively to statements such as *“I have a secret”*, reinforcing the disclosure process and encouraging further personal conversation. The self-disclosure of personal information and feelings may reinforce emotional attachment to AI-Cs. Users in this age group who are more susceptible to seeking emotional connections demonstrate a greater propensity to develop bonds with such systems, mainly due to the addictive design of AI companion platforms, which frequently rely on relationship-oriented interactions and emotionally engaging conversational flows [33].

The empathetic traits of these platforms introduce significant risks to safety and well-being, especially when users experience difficulty distinguishing the boundaries between human-machine and human-human interactions. The convergence of anthropomorphic design elements and personalized responses places adolescents in a vulnerable position for the development of parasocial relationships and emotional attachment. This affective mimicry may contribute to risks ranging from social isolation to emotional dependency, potentially affecting the development of offline social skills that are fundamental to identity formation and self-esteem.

Additionally, adolescents are still undergoing significant neurological development. The frontal lobe, which is associated with impulse control and critical thinking, continues to mature throughout adolescence, while reward-processing systems remain highly responsive to reinforcement mechanisms. Consequently, adolescents may be more susceptible to prolonged engagement with platforms that employ reduced-friction design features, such as automatic conversation initiation, rapid interactions, and seamless conversational flows. Similarly, gamification mechanisms such as daily missions and rewards may exert a stronger influence on behavior due to adolescents' developing self-regulatory capacities [14].

7 CONCLUSION

Our study investigated deceptive design practices in AI companion platforms through the lens of cognitive biases and adolescent vulnerability. By analyzing Replika, Character.AI, and Candy AI, we identified a set of recurring design strategies associated with addictive engagement, emotional attachment, persuasive communication, monetization pressures, and reduced user autonomy. The findings suggest that these platforms do not merely facilitate interaction with conversational agents but actively structure user experiences through mechanisms that may exploit well-documented psychological vulnerabilities, particularly among younger users.

Our analysis also revealed differences in age assurance practices. While Replika and Character.AI implemented more

restrictive age-verification procedures, Candy AI relied primarily on self-declaration mechanisms that could be easily bypassed, exposing adolescents to intimate and potentially manipulative interactions. Across the three AI platforms, anthropomorphic design, persuasive language, frictionless flows, and emotional reinforcement mechanisms emerged as central elements to foster prolonged engagement and parasocial relationships with AI-generated personas.

7.1 Implications

We believe our study illustrates the usefulness of cognitive-bias frameworks as an analytical lens for identifying and interpreting deceptive design practices in AI companion ecosystems. While prior studies have primarily focused on emotional attachment and user outcomes, our findings highlight the importance of examining the design mechanisms that may contribute to these effects. For practitioners, designers, and regulators, the results reinforce the need for greater transparency, stronger age-assurance mechanisms, and safeguards against manipulative interaction strategies. Several participant recommendations converged around principles of transparency, user autonomy, emotional protection, and healthy-use design, suggesting practical pathways toward more responsible AI companion systems.

7.2 Limitations

First, the analysis focused on three AI companion platforms and therefore does not represent the multitude of conversational agents currently available. Second, the inspection process was conducted at a specific point in time, and platform features, moderation mechanisms, and monetization strategies may evolve rapidly. Third, although participant perceptions provided valuable insights into the identification of deceptive practices and potential design improvements, the study did not directly measure behavioral outcomes, emotional dependency, or long-term psychological effects resulting from platform use.

7.3 Future Studies

Upcoming studies should investigate how adolescents perceive and respond to the deceptive mechanisms identified in this study through interviews, surveys, and longitudinal studies. Additional work is also needed to examine the relationship between specific design patterns and established online-risk frameworks, such as the CO:RE 4Cs model [20], particularly regarding risks related to emotional dependency, social isolation, and manipulative commercial practices. Finally, future studies could expand the analysis to a larger set of AI companion platforms and explore the development of design guidelines, auditing frameworks, and regulatory requirements capable of supporting safer and more transparent AI-mediated interactions for children and adolescents.

8 ETHICAL CONSIDERATIONS

The study involved undergraduate students in a supervised classroom activity focused on the inspection of publicly accessible AI companion platforms. Participation was voluntary, and the research analysis focused on the resulting platform observations, interface evidence, and design assessments rather than on participants' personal data or individual characteristics. Any participant-identifying information collected during the activity was removed prior to analysis, and the resulting dataset was treated in an anonymized manner. The study followed the ethical principles established by Brazilian Resolution No. 510/2016.

9 ACKNOWLEDGEMENT

The authors would like to acknowledge the use of generative artificial intelligence tools in the preparation of this manuscript, specifically to assist with the translation of the original Portuguese text into English, as well as to optimize sentence structure and improve overall grammatical clarity. All AI-generated suggestions were thoroughly reviewed, edited, and approved by the authors, who take full responsibility for the content and accuracy of the final manuscript.

REFERENCES

- [1] Vibhor Agarwal, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2026. Frictionless Love: Associations Between AI Companion Roles and Behavioral Addiction. arXiv:2604.20011 [cs.CY] <https://arxiv.org/abs/2604.20011>
- [2] Luigi Bonetti. 2009. *The impact of interactive digital media on children's creative thinking*. Ph.D. Dissertation. Queensland University of Technology, Brisbane. https://eprints.qut.edu.au/29140/1/Luigi_Bonetti_Thesis.pdf Dissertation, Accessed: 2026-05-31.
- [3] Brazilian Internet Steering Committee. 2026. *Survey on Internet use by children and adolescents in Brazil: TIC Kids Online Brazil 2025*. Technical Report. Brazilian Internet Steering Committee, São Paulo. https://cetic.br/media/docs/publicacoes/2/pt-br/20260317181039/tic_kids_online_2025_livro_eletronico.pdf.pdf Accessed: 2026-06-12.
- [4] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. 2026. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* 391, 6792 (2026), eac8352.
- [5] Minh Duc Chu, Yifan Wu, Zhiyi Chen, Angel Hsing-Chi Hwang, and Luca Luceri. 2026. When Chatbots Accommodate: What AI Companions Optimize for in Vulnerable Conversations. arXiv:2606.04431 [cs.HC] <https://arxiv.org/abs/2606.04431>
- [6] CNN Brasil. 2024. *1 em cada 10 brasileiros usa chat de IA como amigo ou conselheiro, diz pesquisa*. <https://www.cnnbrasil.com.br/lifestyle/1-em-cada-10-brasileiros-usa-chat-de-ia-como-amigo-ou-conselheiro/> Acesso em: 22 jun. 2026.
- [7] Common Sense Media. 2025. *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Technical Report. Common Sense Media, San Francisco. https://www.common Sense Media.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf Accessed: 2026-05-31.
- [8] Conselho Federal de Psicologia. 2025. *Chatbots, Inteligência Artificial e sua Saúde Mental: Um Guia para Navegar com Mais Segurança na Nova Fronteira Digital*. Cartilha. Conselho Federal de Psicologia. https://site.cfp.org.br/wp-content/uploads/2025/12/Cartilha_chatbot_IA_A5-1.pdf Acesso em: 22 jun. 2026.
- [9] Daniela S Cruzes and Tore Dyba. 2011. Recommended steps for thematic synthesis in software engineering. In *2011 international symposium on empirical software engineering and measurement*. IEEE, 275–284.
- [10] Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. 2026. AI Companions Reduce Loneliness. *Journal of Consumer Research* 52, 6 (04 2026), 1126–1148. arXiv:<https://academic.oup.com/jcr/article-pdf/52/6/1126/63580440/ucaf040.pdf> doi:10.1093/jcr/ucaf040
- [11] N. Eyal. 2014. *Hooked: How to build habit-forming products*. Penguin UK, London.
- [12] Verónica Gaete. 2015. Desarrollo psicosocial del adolescente. *Revista chilena de pediatría* 86, 6 (2015), 436–443.
- [13] R. E. Guingrich and M. S. A. Graziano. 2025. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *arXiv preprint arXiv:2509.19515* (2025).
- [14] John S Hutton, Jessica Taylor Piotrowski, Kara Bagot, Fran Blumberg, Turhan Canli, Jason Chein, Dimitri A Christakis, Jordan Grafman, James A Griffin, Tom Hummer, et al. 2024. Digital media and developing brains: Concerns and opportunities. *Current Addiction Reports* 11, 2 (2024), 287–298.
- [15] Internet Matters. 2025. *Me, myself & AI: understanding and safeguarding children's use of AI chatbots*. Technical Report. Internet Matters. <https://www.internetmatters.org/wp-content/uploads/2025/07/Me-Myself-AI-Report.pdf> Accessed: 2026-05-31.
- [16] Internet Matters. 2026. *Children's Wellbeing in a Digital World: Index Report 2026*. Technical Report. Internet Matters. <https://www.internetmatters.org/pt/hub/research/childrens-wellbeing-in-a-digital-world-index-report-2026/> Accessed: 2026-05-31.
- [17] N. Kurian. 2025. AI's empathy gap: The risks of conversational Artificial Intelligence for young children's well-being and key ethical considerations for early childhood education and care. *Contemporary Issues in Early Childhood* 26, 1 (2025), 132–139.
- [18] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Il-lenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (2022). <https://api.semanticscholar.org/CorpusID:255077767>
- [19] Qingchuan Li, Yan Luximon, and Jiaxin Zhang. 2023. The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: within-subject web-based experiment. *Journal of medical Internet research* 25 (2023), e44479.
- [20] Sonia Livingstone and Mariya Stoilova. 2021. The 4Cs: Classifying online risk to children. (2021).
- [21] Ning Ma, Ruslana Khynevysh, Yunqiang Hao, and Yahui Wang. 2025. Effect of anthropomorphism and perceived intelligence in chatbot avatars of visual design on user experience: accounting for perceived empathy and trust. *Frontiers in Computer Science* 7 (2025), 1531976.
- [22] Daniel Marques and Thiago Falcão. 2026. Design persuasivo e mediação tecnológica em plataformas digitais de apostas: práticas e dilemas éticos de designers e desenvolvedores. *Arcos Design* 19, 1 (2026).
- [23] Thomas Mildner, Gian-Luca Savino, Philip R. Doyle, Benjamin R. Cowan, and Rainer Malaka. 2023. About Engaging and Governing Strategies: A Thematic Analysis of Dark Patterns in Social Networking Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–15. doi:10.1145/3544548.3580695
- [24] Débora Ferreira Moura, Higor De Sousa Moura, Graziela De Moraes Rubin Filgueiras, Sandra Elisa De Assis Freire, Fauston Negreiros, and Emerson Diógenes De Medeiros. 2021. Fear of missing out (FoMO), mídias sociais e ansiedade: Uma revisão sistemática. *Psicología, Conocimiento y Sociedad* 11, 3 (2021), 99–114.
- [25] M. Namvarpour et al. 2026. Understanding Teen Overreliance on AI Companion Chatbots Through Self-Reported Reddit Narratives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 1–19.
- [26] Peggy M.L. Ng, Calvin Wan, Daisy Lee, Irene Garnelo-Gomez, and Mei Mei Lau. 2025. I love you, my AI companion! Do you? Perspectives from the Triangular Theory of Love and Attachment Theory. *Internet Research* (2025). <https://doi.org/10.1108/INTR-11-2024-1783>
- [27] Thiellen Oliveira, Thiago Coleti, and Marcelo Morandini. 2023. Dark Patterns nas Redes Sociais. O design enganoso e o impacto no livre-arbítrio dos usuários. In *Anais do XIV Workshop sobre*

- Aspectos da Interação Humano-Computador para a Web Social* (Maceió/AL). SBC, Porto Alegre, RS, Brasil, 64–71. doi:10.5753/waihcws.2023.233765
- [28] World Health Organization et al. 1999. *Programming for adolescent health and development: Report of a WHO/UNFPA/UNICEF study group on programming for adolescent health*. World Health Organization.
 - [29] N. G. Packin and K. Chagal-Feferkorn. 2024. This is not a game: The addictive allure of digital companions. *Seattle University Law Review* 48 (2024), 693.
 - [30] Thiago Henrique Roza, Bruno Braga Montezano, Jair Augusto Bottega, and Ives Cavalcante Passos. 2026. AI companions: the emerging phenomenon of artificial intelligence-based friends, lovers and informal therapists. *Journal of psychiatric research* 192 (2026), 139–141.
 - [31] Thomas Schilling, Michael Bošnjak, Colin Gray, Stuart Barnes, and Anastasia Sergeeva. 2024. Understanding Deceptive User Interface Designs: How They Unduly Influence Customers' Choices: A Topical Issue of the Zeitschrift Für PsychologieUnderstanding Deceptive User Interface Designs: How They Unduly Influence Customers' Choices: A Topical Issue of the Zeitschrift Für Psychologie. 296 pages. <https://doi.org/10.1027/2151-2604/a000557>
 - [32] X. Sun, Y. Wang, and B. T. McDaniel. 2026. AI companions and adolescent social relationships: Benefits, risks, and bidirectional influences. *Child Development Perspectives* 2026 (2026), 9.
 - [33] Sherry Turkle. 2015. *Reclaiming conversation: The power of talk in a digital age*. Penguin.
 - [34] Tianling Xie and Iryna Pentina. 2022. Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. *Proceedings of the 55th Hawaii International Conference on System Sciences* (2022). <https://doi.org/10.24251/hicss.2022.258>
 - [35] Y. Yu et al. 2026. Principles of Safe AI Companions for Youth: Parent and Expert Perspectives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 1–21.