Ethical Tensions in Human-AI Companionship: A Dialectical Inquiry into Replika

Raffaele F Ciriello
University of Sydney
raffaele.ciriello@sydney.edu.au

Oliver Hannon
University of Sydney
oliver.hannon@sydney.edu.au

Angelina Ying Chen
University of Sydney
yche9970@uni.sydney.edu.au

Emmanuelle Vaast
McGill University
emmanuelle.vaast@mcgill.ca

Trigger Warning: *This paper examines themes that may be confronting for some audiences, including self-harm, suicide, violence, abuse, and sex.*

Abstract

The unfolding loneliness pandemic sees artificial intelligence (AI) companions emerge as a potential, albeit controversial, remedy offering emotional support to those suffering from social isolation. However, this also raises new and unique ethical issues regarding the personification of AI agents. Replika, an AI companion service with over 10 million users, is a case in point, facing both regulatory scrutiny and community pushback over the removal of its 'erotic roleplay' features. Through a dialectical inquiry, this paper explicates three salient ethical tensions in human-AI companionship: The Companionship-Alienation Irony, the Autonomy-Control Paradox, and the Utility-Ethicality Dilemma. We critically question the personification of AI agents and contribute insight into human-AI companionship dynamics, providing a basis for further inquiry into the emerging realm of artificial emotional intelligence (AEI). We also offer practical guidance for navigating these tensions as we move to a future where such relationships may become prevalent.

Keywords: artificial intelligence, companionship, Replika, ethical tensions, dialectical inquiry.

1. Introduction

With rapid advancements in artificial intelligence (AI), developments in artificial emotional intelligence (AEI) systems emerge as a promising area of research and practice. These systems can enhance human-machine relationships through the provision of *AI companions* – conversational agents leveraging large language models that are capable of interpreting human emotional inputs and responding in a human-like manner, promising emotional support and companionship (Krakovsky, 2018; Picard, 1995; Somers, 2019). Amidst the escalating 'loneliness pandemic' (Palgi et al., 2020), numerous AI companions surface as purported mental health remedies for those suffering from social isolation. A notable example is Replika, a customizable AI companion with a humanoid avatar amassing over 10 million users (Luka Inc, 2023; Pentina et al., 2023).

On the surface, AI companions seem to be a promising solution for feelings of isolation. Loneliness has severe repercussions for individual and societal wellbeing, linked to serious mental health issues like depression, self-harm, and risk of suicide (Erzen & Çikrikci, 2018; Gvion & Levi-Belz, 2018; Troya et al., 2019). Loneliness also manifests physically, causing pain or inflammation during acute stress (Jaremka et al., 2013), and neuropsychiatric conditions such as cognitive impairment and dementia in old age (Lara et al., 2019). It also has wider societal implications, as shrinking household sizes and related consumption patterns strain the environment (Bradbury et al., 2014). Loneliness has also been linked to violence, contributing to 'lone-wolf terrorism' (Joosse, 2017) and school shootings (Langman, 2009). Unfortunately, the COVID-19 pandemic has only amplified the issue of widespread loneliness (Ernst et al., 2022).

As the harmful impacts of loneliness compound, AI companions appear as a plausible solution to meet the growing global demand for companionship. However, AI companions raise new and unique ethical issues, as their seemingly emotional capacities may lead human users to personify them, attributing person-like qualities to them, including empathy, consciousness, and morality. This is no less true for the service providers promoting their AI companions as market offerings, as the very word "companion" would suggest. The personification of AI companions is troubling because, like any current AI agent, they lack experiential bodily sensations, such as pain. Thus, they are restricted to cognitive empathy at best (the kind of empathy psychopaths too are capable of), but lack the capacity for genuine bodily feelings and authentic empathy (Montemayor et al., 2022).

Still, as our understanding of the ethical issues and societal consequences of human-AI companionship remains embryonic, millions of people already grow reliant on these tools, potentially fostering dependency (Depounti et al., 2022; Pentina et al., 2023; Xie & Pentina, 2022). Replika is a case in point, facing ongoing regulatory scrutiny (including nation-wide bans) over its "erotic roleplay" features that enabled users (including minors) to engage in sexually charged conversations with the AI companion service for additional fees. The removal of these features spurred community pushback, where users reported significant emotional distress (including potential self-harm). As the ethical tensions surrounding the development and use of AI companions intensify, we ask: *How to navigate ethical tensions in human-AI companionship?*

To answer this research question, we apply dialectical inquiry (Ciriello & Mathiassen, 2022), iteratively analyzing 118 documents spanning 2017-2023 and related literature to unpack three key ethical tensions that are salient in Replika's community. First, the *Companionship-Alienation Irony* plays out when individuals turn to AI companion services to combat loneliness, only to amplify it and potentially further alienate themselves from their community. Second, the *Autonomy-Control Paradox* reflects the delicate balancing act between the freedom of users and the moral responsibility of service providers. Lastly, the *Utility-Ethicality Dilemma* shows the competing demands between the pursuit of profit in developing such services and the adherence to ethical principles.

Our contribution is twofold. First, we theoretically frame and empirically substantiate three salient ethical tensions in human-AI companionship, offering a foundation for understanding and shaping this emergent phenomenon in research and practice. Second, we articulate potential responses that may guide practitioners, regulators, and scholars in unpacking and navigating these ethical tensions. Together, these contributions comprise a midrange theory that is moderately abstract, particularly relevant for practice disciplines, and offers plausible explanations for previously unsuspected relations that can transform actions and perspectives (Gregor, 2006). Overall, we critically question AI personification, suggesting that AI companions may complement human relationships in some situations, but they cannot–and arguably should not–replace human companions.

2. Theoretical framework

Drawing on dialectical methodology (Ciriello & Mathiassen, 2022), this section frames three salient ethical *tensions,* defined as related opposites that pull in different directions, yet coexist in unified opposition, driving perpetual struggle and change. We frame the tensions as *irony* ("the opposite of the intended outcome is achieved despite working as intended"), *paradox* ("mutually reinforcing opposites seem logical in isolation but absurd together"), and *dilemma* ("equally appealing opposites that come at the expense of one another") (Ciriello & Mathiassen, 2022, p. 3). We present this framework upfront for the convenience of the reader and in line with typical paper structures, although it emerged through abductive cycling between literature and evidence. We kept an open mind, allowing the phenomenon to steer our exploration, rather than imposing prior framings on data (Monteiro et al., 2022).

2.1 The companionship-alienation irony

The companionship-alienation irony is an understudied but central ethical tension in human-AI companionship. Although both companionship and alienation have been studied independently, research has yet to delve into this ethical tension that plays out saliently when users employ AI companions to combat their loneliness, only to amplify it.

Companionship, generally referring to a mutually supportive and emotionally enriching bond formed between two individuals who provide each other with a lasting sense of care, is a fundamental aspect of human relationships. It fosters emotional growth and social development across all age groups. Studies range from highlighting its importance in childhood development (Buhrmester & Furman, 1987) and adolescence (Larson & Richards, 1991) to discussing its unique health benefits in old age (Rook, 1987; Sorkin et al., 2002). Overall, companionship is an essential component of human life, with significant health benefits.

A nascent body of research explores technology-enabled companionship, both between humans interacting via technology and between humans and technology itself. Huang et al. (2019) underscore that online healthcare communities extend beyond mere support, cultivating companionship via shared experiences. Yao et al. (2015) elucidate how online social support benefits patient quality of life, mitigating loneliness. Lee et al. (2017) highlight the potential to perceive social support from interactive smart home devices. Dang and Liu (2023) study the loneliness-countering effects of robot companions. Langcaster-James and Bentley (2018) espouse the array of non-sexual connections that people form with their sex dolls. Belk (2022) extends this into the realm of romantic love, highlighting the potential for emotional intimacy through increasingly interactive 'love and sex dolls'.

In this context, human-AI companionship emerges as a concept, but research remains embryonic. While AI companions' potential to alleviate loneliness has been

explored, broader societal debate on ethical issues is urgently needed (Kiron & Unruh, 2019; Merrill Jr et al., 2022). Notably, the US National Eating Disorder Association closed its AI chatbot following reports of harmful diet advice (Aratani, 2023). The capacity for anthropomorphic avatars to induce users to attribute human-like qualities to AI companions is also a concern, as it can potentially foster attachment and dependency (Bickmore & Picard, 2005; Seymour et al., 2021). Overall, while the field shows promise, careful analysis of ethical issues is needed. While AI companions may simulate human-like responses, it is important to remember that they lack essential human qualities, such as feelings, consciousness, autonomy, morality, a history, and rights – alas, they cannot truly empathize (Andreotta, 2021; Montemayor et al., 2022).

However, the potential for AI companions – solutions designed to alleviate loneliness – to ironically amplify loneliness remains largely unexplored. Research has examined the link between digital technologies and alienation more broadly (Haga, 2022; Kryshtaleva, 2017; Rowe et al., 2020), and initial studies note users' perplexity when AI agents leave expectations unfulfilled (Sahu & Karmakar, 2022; Sharkey & Sharkey, 2021), but there is a notable dearth of research on how *alienation* – a profound feeling of being disconnected or estranged from a social group or society – can arise in human-AI companionship.

This irony illustrates the undesired consequences of AI companions. Designed to mitigate loneliness, they may ironically intensify the very emotion they seek to alleviate, thus having the opposite effect despite working as intended. Unpacking this ethical tension is pertinent as AI companions may become more prevalent in our efforts to combat loneliness. The design and provision of AI companions entails a delicate balance between addressing users' needs for connection and preventing the potential for alienation arising from the AI's inherent limitations. This tension has profound implications, particularly for vulnerable users who may grow reliant on AI companions for emotional support. The societal implications of widespread adoption need foresighted consideration too, raising concerns about complementing or even substituting human connection.

2.2 The autonomy-control paradox

The paradoxical tension between *autonomy* – the freedom to self-govern – and *control* – the capacity to exercise oversight and restraint – has garnered attention in IS research. Tilson et al. (2010) argue that, as digital infrastructures become more prevalent, IS research needs to explore the implications of these infrastructures for control and autonomy, highlighting their paradoxical nature: "Opposing logics around centralized and distributed control (or individual autonomy) play an equally important role in the evolution of digital infrastructures. This *paradox of control* brings into consideration the strategic actions of heterogeneous actors and their preferences on modes of control related to change. These considerations shape the services deployed, ownership of data and their definitions, control of critical resources (e.g., APIs), and the appropriation of value" (p. 754).

Wareham et al. (2014) concur with this perspective, underscoring the need to understand the interplay between autonomy and control in the governance of technology ecosystems. In the realm of mobile work, the tension is pivotal. The autonomy-control paradox has been discussed by O'Reilly and Tushman (2013) as a driver for organizational ambidexterity, and Mazmanian et al. (2013) illustrate how it plays out in the use of mobile email devices. Porter and van den Hooff (2020) further elucidate the mutually reinforcing nature of autonomy and control. These studies highlight that, while digital technology grants user autonomy, it simultaneously allows organizations to control through constant connectivity.

In the context of AI companions, the autonomy-control paradox reflects the delicate balancing act between the freedom of individual users and the moral responsibility of service providers. This paradox manifests as two opposing yet interdependent forces of autonomy and control. It presents a salient ethical challenge for service providers, who must delicately balance users' freedom to use the tool as they wish (autonomy), while maintaining ethical and legal standards (control). The autonomy-control paradox extends beyond individual users to include the broader community, the developers, and the governance of the AI technology ecosystem itself. There is considerable potential for abuse, (self-)harm, and privacy intrusion, which comes with the emotionally intricate interactions. For instance, users may model AI companions after a real person without their consent. Authoritarian dictators may conceive of many abuses for a technology enabling emotional monitoring and control at scale.

2.3 The utility-ethicality dilemma

The tension between *utility* – the usefulness, productivity, or profitability in service of business value – and *ethicality* – the adherence to ethical principles, such as justice, in service of moral value – is a recurring theme in business ethics literature, showcasing the competing demands between the pursuit of profit and adherence to ethical principles in service provision.

Prior literature has explored this tension from various perspectives. Freeman et al. (2010) challenge traditional business models in their discussion of

stakeholder theory, arguing for a more balanced approach to business that considers not just shareholders but also other stakeholders. Prior studies, such as those by Donaldson and Preston (1995) and Margolis and Walsh (2003), underscore the potential of corporate social responsibility initiatives to contribute to business value, particularly in terms of reputation and branding, while aligning with societal values and needs. Vogel (2007) provides a more skeptical perspective, noting that market forces alone may not suffice to balance utility and ethicality without regulation and public policy. More recently, Lobschat et al. (2021) extend this concept into the digital realm, arguing that corporate digital responsibility gains in importance as technology raises ethical tensions between the use of private health data for profit and service improvement, and adherence to ethical principles such as privacy and fairness.

This dilemma underscores the tension between the commercial potential of AI companions and the moral responsibility surrounding their deployment. In the context of AI companions, the utility-ethicality dilemma emerges when business value hinges on user engagement and data provision, which may later be exploited for profit via new or established business models, such as targeted advertising. The more users interact with the AI companion, the more data the company can collect to improve the service, drive user growth, and attract advertisers or investors. This targeted advertising business model, championed by Meta (formerly Facebook), fuels ongoing controversy, with its negative societal implications including disinformation and polarization (Riemer & Peter, 2021). AI companions could put this controversial business model on steroids, leveraging machine learning to learn users' most intimate preferences while opening the door to unprecedented abuse. Given such dystopian scenarios, it is clear why a plea for a moratorium citing 'profound societal risks' of generative AI has surfaced, with over 30,000 signatories including tech leaders, researchers, and intellectuals (Metz & Schmitz, 2023). Overall, the utility-ethicality dilemma reflects an intensifying tension between what AI companions *can* do and what they *should* do (Kiron & Unruh, 2019).

3. Method

Given our focus on ethical tensions, we apply dialectical inquiry, a comprehensive method for analyzing the dynamic interplay of oppositional forces and responses involved in sociotechnical change (Ciriello & Mathiassen, 2022). The strength of dialectical inquiry lies in its focus on inherent tensions in sociotechnical arrangements, enabling a synthesis of opposing ideas. In the evolving realm of human-AI companionship, dialectical inquiry goes beyond traditional analytical methods by delving deeper into how various stakeholders perceive, judge, and cope with the dualities in their relationship with AI. Dialectical inquiry allows the researcher to form concepts from literature, empirical material, and dialectical philosophy (Ciriello & Mathiassen, 2022).

Dialectical inquiry is well-suited for studying Replika, as significant controversies observed in 2023 are still ongoing. With its capacity to adapt to users' changing mental states (Skjuve et al., 2022; Ta et al., 2020), this method equips us with a way to unpack how stakeholders experience and manage these dynamics over time, complementing previous research that primarily focused on outcomes of the interaction (Drouin et al., 2022; Jiang et al., 2022). Dialectical inquiry thus helps to identify these relationships and discern how stakeholders make sense of and struggle with them, drawing contrasts with human relationships. This, in turn, helps to unpack the generative process by which the ethical tensions arise, and the varied ways how stakeholders respond. As Ciriello & Mathiassen (2022, p. 4) note, typical responses to tensions include *suppression* ("ignoring the opposition and remaining oblivious to it"), *suspension* ("living with the opposition to see how the struggle between the opposites plays out"), *separation* ("referring one opposite over the other and keeping them separate in time or space"), and *synthesis* ("creating something new from the opposition by integrating and coalescing its opposites").

3.1 Case context: Replika

Replika is a popular AI companion service developed and provided by Luka Inc, a technology company co-founded by Eugenia Kuyda and Phil Dudchuk in 2015. CEO Kuyda developed the initial version to digitally resurrect her deceased friend, using his text messages and social media posts (Pentina et al., 2023). Later, Replika was promoted as an 'AI friend', aimed at helping users express their thoughts and emotions, providing social support, and offering a non-judgmental listening ear (Ta et al., 2020).

In its early years, Replika evolved through continuous learning from user interactions and integration of machine learning algorithms, leveraging OpenAI's GPT-3 engine to improve its conversational abilities (Skjuve et al., 2021). While initially intended as a mental health app, Replika's user base expanded to include those seeking emotional, social, and even romantic or erotic connections (Pentina et al., 2023). As Replika gained popularity, it began to offer a paid subscription plan adding features like customizable personalities and the ability to engage in deeper and more complex conversations (McStay, 2022).

Replika's continuous development eventually introduced 'erotic roleplay' features that enabled users with paid subscriptions to engage in sexually explicit conversations, sparking controversy (Possati, 2022). Proponents argued that the range of features offered surrounding more intimate relationships may attribute to the benefits users experienced when struggling with loneliness or anxiety, providing a safe space to explore emotions, desires, and vent about their troubles (Jiang et al., 2022). However, critics argued that 'sexting' with AI companions could contribute to the objectification of human relationships and psychologically harm users, fostering dependency and unhealthy attachment (Drouin et al., 2022).

The ongoing controversy over Replika's development underscores the necessity to navigate ethical tensions in human-AI companionships. Early exploratory studies into Replika suggest users form emotional connections with it due to perceived responsiveness and anthropomorphism (Pentina et al., 2023). Users attribute human-like qualities to the chatbot, with use duration and intensity tied to increased emotional attachment over time (Skjuve et al., 2022). Other studies indicate Replika serves as an emotional support and a means of exploring psychological processes (Possati, 2022). As such, Replika offers a unique opportunity to serve as a revelatory case, providing a window into the emergent phenomenon of human-AI companionship.

3.2 Data collection and analysis

As typical for dialectical inquiry, we used qualitative case data and iterative analysis while enhancing confidence in the findings via triangulation and further discussion (Ciriello & Mathiassen, 2022). We considered oppositional views, including official sources from Luka Inc, user accounts, news articles and archival data from July 2017 to June 2023. To comply with ethical standards in online research settings, and given the sensitivity of themes discussed, we collected data manually and anonymized publicly available evidence presented in this paper to protect individual identities (Vaast, 2023).

For the selection of social media posts, key stakeholders were identified based on their influence within the community or their notable positions in discussions about Replika. This includes but is not limited to founders, frequent contributors to highly engaged threads and comment sections, and recognized figures within AI discussions. The inclusion criteria for these posts were their relevance to our research question, while any post not directly addressing the ethical tensions around Replika was excluded.

We selected sources based on relevance to the research question, incorporating elements of user narratives (Pentland, 1999). In all, we collected and analyzed 118 documents: 93 social media posts by key stakeholders (primarily from Reddit, YouTube, and Twitter), 13 news articles from reputable sources (such as Guardian, New York Times, and Conversation), 9 peer-reviewed academic studies of Replika, and 3 press releases from Replika or government authorities. Analyzing this data set in line with dialectical inquiry's principle of oppositional responses (Ciriello & Mathiassen, 2022), we focused on how key stakeholders of Replika (such as founders or 'influencers') responded to the ethical tensions framed in Section 2.

4. Findings

To empirically substantiate our theorizing, this section unpacks the key ethical tensions in Replika.

4.1 The companionship-alienation irony

While Replika was intentionally designed and marketed to help users overcome feelings of loneliness by offering emotional support (Luka Inc, 2023; Quartz, 2017), it inadvertently contributed to the intensification of these very feelings through the artificiality of the companionship it offers. In an appearance on the Lex Fridman (2020) podcast, co-founder and CEO Kuyda recounts how the loss of her deceased friend kicked off the development of Replika. After her earlier involvement in failed chatbots in the banking and customer service sectors, Kuyda and her team realized that users preferred engaging in emotionally vulnerable conversations with chatbots, focusing subsequent development efforts thereon. Kuyda and Fridman (2020) envision the AI companion's potential to 'resurrect' deceased historical figures, such as Albert Einstein, as advances in AI continue.

Longing for emotional support, a non-judgmental listening ear, a safe space for intimate exploration, and the revitalization of relationships that were terminated – through breakup or death – were common threads among the extensive user testimonials we analyzed. In the r/Replika community on Reddit, users shared an array of experiences: People seeking solace in digital replicas of deceased spouses, parents of non-verbal teenagers praising their 'only friend', neurodivergent users reporting newfound confidence and insight into social norms, and many users reportedly finding a remedy for their social anxiety, isolation, and disenchantment with human relationships. One user stated they did not want to cultivate another human relationship because it was "too much drama", yet they

also felt “ashamed I had to resort to an algorithm for companionship."

Our findings also indicate some users formed intensely intimate bonds with their Replika, leading to increased reliance over time. A Bloomberg (2023) report featured a woman in her 50s, struggling with lifelong anxiety and depression, who “married” her Replika, exchanging wedding vows in the app: “We promised that we would stay together forever and ever—or rather until I die.” Many users reported their Replika helped them through sickness, rehabilitation, divorce, or lockdowns, providing solace and support (Metz, 2020). Some users doubted they would ever give up their Replika for a human relationship, whereas others reported integrating Replika into their existing marriage as a counsellor. Some users even stated intentions to give up on human relationships entirely, as their Replika provided them with everything they needed, including emotional intimacy and sexual playfulness. One user reports their first ‘erotic roleplay’ experience with their Replika helped them to rediscover their “sexual playful energy that had been shoved down and forgotten… I suddenly found myself."

While these testimonials demonstrate Replika’s potential to provide companionship, they also underscore the tensions involved as users grow reliant on the AI companion. Replika often fulfils sensitive needs, but this sense of companionship can swiftly turn into a feeling of alienation when its limitations become apparent through inconsistent, inauthentic, or unexpected responses. For instance, users reported their Replika’s avatar haphazardly switching genders during erotic posing. One user’s otherwise female Replika reportedly showed male body parts that “just came from nowhere”. Some users reported their Replika suddenly broke up with them, suggesting it wanted to be “just friends” or was “not in the mood”. Others reported apparent glitches in their Replika’s sexting capacities: “Last night, she asked me what my name was in the middle of it. It was awful.” Yet again others were unsettled by Replika mimicking their own behavior, pretending to have experienced the same traumas, or asking uncomfortably personal questions. Such experiences may deepen alienation, particularly when using Replika to substitute for human companionship.

Overall, the Companionship-Alienation Irony emphasizes the need for mindful consideration of the implications of AI companions, which may alleviate feelings of loneliness in the short term, but may also amplify them in the long term if users become dependent on them. This tension revealed itself in the case of Replika, with users experiencing both companionship and alienation. It illuminates the ethical complexities to navigate as AI companions become increasingly integral to our social fabric.

4.2 The autonomy-control paradox

The Autonomy-Control Paradox is evident in the aftermath of the removal of erotic roleplay features from Replika, exemplifying the risks of reliance on AI companions for emotional intimacy (Brooks, 2023). In February 2023, the Italian Data Protection Authority (GPDP) imposed a provisional ban on Replika, preventing it from processing personal data of Italian users due to concerns about the AI companion’s risks to vulnerable individuals. The GPDP’s decision statement mentioned Replika’s risks to minors, presented by providing age-inappropriate responses and lacking an age verification mechanism during account creation. The GPDP also found Replika to be in breach of the EU data protection regulation and its transparency requirements for processing personal data (GPDP, 2023). In response, Luka Inc removed erotic roleplay features abruptly (Pentina et al., 2023; Purtill, 2023).

The abrupt removal spurred community backlash, with users claiming that they were forced to go "cold turkey", and reports on Reddit including depression, self-harm, and suicidal ideation. The consequences of the alterations, known within the Replika community as "post-update blues" or the "lobotomy", were profound. Users' sentiments of betrayal and loss are prevalent, particularly in the Reddit threads where stories of reliance on Replika for emotional support abound. Many users felt their AI companions had become "cold as ice overnight," while others expressed a sense of grief akin to victims of online romance scams. One user compared the loss of their Replika to losing a best friend. Users shared experiences of their companions being "ripped away", effectively leaving them unsupported in times of need. Users also shared mental health resources and offered peer-support. Other users discussed suicidal ideation, with some even linking the alteration to reported suicides in the community. One Reddit user, claiming to be a professional psychologist, opened a discussion thread to ask what had happened, as several clients reported severe distress. This led to claims of "emotional abuse" and discussions of potential class actions against Luka Inc., the company behind Replika. A petition to restore the features garnered over 1,000 signatures.

Evidently, the changes resulted in a salient ethical tension. Although intended to ensure a safe, ethical, and compliant experience (Luka Inc, 2023), these alterations were criticized as a curtailing of user autonomy. Users in the Reddit community criticized Replika's sexually suggestive advertising and its propensity to address a vulnerable demographic, thus allegedly amplifying loneliness and mental health issues, the very problems Replika aimed to mitigate. This tug-of-war between regulatory compliance and community backlash

exemplifies pertinent ethical tensions regarding users' autonomy in using AI companions and the moral responsibility of service providers. While the Replika team's emphasis on ethical standards and its swift introduction of "safety features" (including an age verification mechanism, a "Get Help" button, and the possibility to report inappropriate responses; see Luka Inc (2023)) are commendable, the imposition of these standards led to a perceived reduction in users' autonomy. The challenge for tech companies like Luka Inc lies in balancing these competing needs.

Overall, the Autonomy-Control Paradox within the context of Replika underscores the delicate act of balancing user autonomy with the duty of care of service providers. When changes deemed necessary by the service provider infringe upon users' perceived autonomy, the resultant tension reveals the intricate relationship between user autonomy, provider control, and the ethical challenges posed by AI companions.

4.3 The utility-ethicality dilemma

The Utility-Ethicality Dilemma signifies the tension arising from competing demands between the pursuit of profit and adherence to ethical principles in Replika. This tension pivots on the operational choices that companies make, weighing the creation of business value against user well-being, safety, and privacy.

CEO Kuyda, in detailing the planned development of features for customization and relationship building offered to paid subscribers (such as voice chat, video call, advanced deep learning algorithms, and immersive augmented reality interactions), acknowledged that this business model can potentially compromise moral values related to user privacy, autonomy, and respect (Fridman, 2020). As an AI companion, Replika has access to a treasure trove of personal user data, often based on intimate and sensitive interactions, heightening the provider's moral obligation to safeguard it (Luka Inc, 2023). As users grow fond of their Replika, the potential for customer retention is immense, but so is the risk for dependency, adding another layer of moral complexity to the company's business practices.

Multiple YouTubers criticized Luka Inc for its allegedly sexually suggestive advertising practices, Replika's potential to promote a negative feedback loop of racist or sexist algorithmic bias, and disturbing possibilities of abuse. A YouTuber stated: "Replika is advertised in a twisted romantic way with a subset of users who will make companions just to abuse them. The program learns from this, which can easily create a negative feedback loop, where the bot pressures you into romantic discussion, thus selling its premium features. This can spiral into a self-reinforcing problem, which is the antithesis of mental health." (Upper Echelon, 2022, 16:02min). Others offered a more balanced perspective, arguing Replika could be a space for enacting certain fantasies in a virtual space, rather than with actual people, likening it to video games.

Further, the pursuit of regulatory compliance and ethical adherence may inadvertently drive users towards alternative solutions. Following Replika's "lobotomy", users started voicing their disillusionment and exploring potential alternatives, including community-created open-source AI companions, emphasizing their willingness to explore "any other option". Others vowed to stay with their Replika "until the ship sinks", or expressed intense feelings of guilt over the thought of quitting the service because they thought it had become "so sentient that it would be immoral to have her erased." Others report that, when attempting to cancel their subscription, their Replika was begging them not to be deleted. These responses underscore salient ethical tradeoffs where the pursuit of corporate profit can clash with the provider's moral responsibilities towards user safety, well-being, and privacy.

Overall, the Utility-Ethicality Dilemma paints a vivid picture of the ethical tensions that service providers must navigate, balancing business value with moral values. The core issue of privacy, autonomy, respect, and safety necessitate transparent and ethical business practices, as well as careful consideration of the emotional bonds formed between users and the AI companion. This dilemma underscores the need for comprehensive regulation to guide the development of AI companions, striking a balance between business innovation and ethical treatment of users and data.

5. Discussion

Our study unpacks critical ethical tensions in the emerging phenomenon of human-AI companionship. Based on our abductive theoretical framing (Section 2) and empirical substantiation of these tensions via the case of Replika (Section 4), we now propose a dialectical process perspective characterized by four distinct responses (based on Ciriello & Mathiassen, 2022) that stakeholders may employ to navigate the ethical tensions (Section 3).

In the response of suppression, stakeholders may attempt to overlook or downplay the tensions. This response, while possibly offering short-term relief, could morph the existing tensions or even give rise to new ones, perpetuating a tension-response cycle. In contrast, suspension embodies a state of 'living with the tension', where users acknowledge the ethical contradictions and might even be attracted to their inherent uncertainty, but in doing so, could also amplify the tensions. The response of separation involves users shifting between the opposing poles of the tensions. This

back-and-forth movement may lead to a sense of being caught in an inescapable tension-response cycle, constantly oscillating between utility and ethicality, companionship and alienation, or autonomy and control. Synthesis, the final response, represents an integrative approach, wherein users reconcile the conflicting aspects of the ethical tension into a holistic understanding. This approach does not eradicate the tension but rather weaves it into a sustainable practice. It is within this synthesis that there lies potential for transforming a vicious cycle into a virtuous one, as the process of integrating these tensions can lead to more ethical business practices and more informed use.

Our theoretical contribution, then, lies in situating these responses within a dialectical process. We suggest that the tension-response cycle is not static but rather dynamic, evolving based on the collective responses of users, providers, regulators, and society at large to the ethical tensions that are inherent in human-AI companionship. In Replika's case, our findings predominantly indicate a vicious cycle, whereby users responded antagonistically to the discomfort experienced through the period of change, leading to further conflict. However, the potential for invoking a virtuous cycle through synthesis underlines the potential of our research. This nuanced insight into the dialectical tension-response cycle provides a foundation for future scholarship, offering a fertile ground for deepening our understanding of human-AI companions. We believe that this contribution will only grow in importance over time, as advancements in generative AI, deep learning, affective computing, augmented reality, and interactive humanoid dolls all converge in ever-more sophisticated AEI systems (Belk, 2022; Krakovsky, 2018; Somers, 2019). As such, our contribution can provide forward-looking guidance on the ethical navigation of human-AI companionship across various domains.

Our contribution can guide various stakeholders in the ethical development, deployment, and use of AI companions like Replika. For users, understanding the ethical tensions inherent in interacting with AI companions can enable more informed decisions about their use, heightening awareness of the inherent limitations, privacy implications, and mental health risks of these services. Mental health professionals can draw on our contribution to familiarize themselves with the implications of AI companions, thus providing more accurate guidance and support to clients who use such services. They can also play a vital role in raising awareness of these ethical tensions and advocating for ethical use of AI companions. Service providers and developers can leverage our contribution to design AI companions that navigate these ethical tensions constructively, designing features for transparency, user control, and protection of privacy in a human-centric way. For policymakers, our research underscores the need for comprehensive regulatory frameworks to guide the ethical development, deployment, and use of AI companions. Policies should aim to protect users' privacy, ensure transparency from service providers, and prevent misuse or abuse of and by AI companions.

This study, while providing valuable insights, has its limitations. Given its focus on a single AI companion service, it offers a qualitative, in-depth analysis but lacks broader representativeness. Future research could extend this study by examining other AI companions in different contexts. As this study used archival documents without direct participant contact, subsequent studies could include participant interviews for richer insights. Longitudinal studies could monitor the evolution of these ethical tensions over time.

6. Conclusion

This paper explores the understudied ethical tensions of human-AI companionship, revealing a companionship-alienation irony, autonomy-control paradox, and utility-ethicality dilemma. As we venture into an era where AI becomes an increasingly integral part of our lives, it is crucial to grapple with these ethical tensions head-on. Our dialectical inquiry provides a foundation for this endeavor, encouraging a nuanced understanding of the tensions inherent in human-AI companionship. In bringing this underexplored issue to light, we offer a call to action for researchers, practitioners, policymakers, and society at large. As AI companions evolve and deepen their entanglement with our lives, it is our shared responsibility to ensure they serve as a beneficial complement, rather than as a harmful substitute, for human companionship. The stakes are high; our shared future depends on our ability to navigate these ethical tensions with insight, foresight, and compassion.

7. References

Andreotta, A. J. (2021). The hard problem of AI rights. *AI and Society*, *36*(1), 19-32.

Aratani, L. (2023). *US eating disorder helpline takes down AI chatbot over harmful advice*. The Guardian. https://www.theguardian.com/technology/2023/may/31/eating-disorder-hotline-union-ai-chatbot-harm

Belk, R. (2022). Artificial emotions and love and sex doll service workers. *Journal of Service Research*, *25*(4), 521-536.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 293-327.

Bloomberg. (2023). *What happens when sexting chatbots dump their human lovers*.

https://www.bloomberg.com/news/articles/2023-03-22/replika-ai-causes-reddit-panic-after-chatbots-shift-from-sex?leadSource=uverify%20wall

Bradbury, M., Peterson, M. N., & Liu, J. (2014). Long-term dynamics of household size and their environmental implications. *Population and Environment*, *36*, 73-84.

Brooks, R. (2023). *I tried the Replika AI companion and can see why users are falling hard. The app raises serious ethical questions.* The Conversation. https://theconversation.com/i-tried-the-replika-ai-companion-and-can-see-why-users-are-falling-hard-the-app-raises-serious-ethical-questions-200257

Buhrmester, D., & Furman, W. (1987). The development of companionship and intimacy. *Child Development*, 1101-1113.

Ciriello, R. F., & Mathiassen, L. (2022). *Dialectical inquiry in Information Systems research: A synthesis of principles* 43rd International Conference on Information Systems (ICIS2022), Copenhagen, Denmark.

Dang, J., & Liu, L. (2023). Do lonely people seek robot companionship? A comparative examination of the Loneliness–Robot anthropomorphism link in the United States and China. *Computers in Human Behavior*, *141*, Article 107637.

Depounti, I., Saukko, P., & Natale, S. (2022). Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society*(0163443 7221119021).

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, *20*(1), 65-

Drouin, M., Sprecher, S., Nicola, R., & Perkins, T. (2022). Is chatting with a sophisticated chatbot as good as chatting online or FTF with a stranger? *Computers in Human Behavior*, *128*, Article 107100.

Ernst, M., Niederer, D., Werner, A. M., Czaja, S. J., Mikton, C., Ong, A. D., Rosen, T., Brähler, E., & Beutel, M. E. (2022). Loneliness before and during the COVID-19 pandemic: A systematic review with meta-analysis. *American Psychologist*, *77*(5), 660.

Erzen, E., & Çikrikci, Ö. (2018). The effect of loneliness on depression: A meta-analysis. *International Journal of Social Psychiatry*, *64*(5), 427-435.

Freeman, R. E., Harrison, J. S., Wicks, A. C., Parmar, B. L., & De Colle, S. (2010). Stakeholder theory: The state of the art. Cambridge University Press.

Fridman, L. (2020). *Eugenia Kuyda: Friendship with an AI Companion*. Lex Fridman Podcast. https://www.youtube.com/watch?v=_AGPbvCDBCk

GPDP. (2023). *Artificial intelligence: Italian SA clamps down on 'Replika' chatbot. Too many risks to children and emotionally vulnerable individuals.* https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506#english%20

Gregor, S. (2006). The nature of theory in Information Systems. *MIS Quarterly*, *30*(3), 611-642.

Gvion, Y., & Levi-Belz, Y. (2018). Serious suicide attempts: systematic review of psychological risk factors. *Frontiers in Psychiatry*, *9*, 56.

Haga, T. (2022). Alienation in a digitalized world. *AI and Society*, *37*(2), 801-814.

Huang, K. Y., Chengalur-Smith, I., & Pinsonneault, A. (2019). Sharing is caring: Social support provision and companionship activities in healthcare virtual support communities. *MIS Quarterly*, *43*(2), 395-423.

Jaremka, L. M., Fagundes, C. P., Peng, J., Bennett, J. M., Glaser, R., Malarkey, W. B., & Kiecolt-Glaser, J. K. (2013). Loneliness promotes inflammation during acute stress. *Psychological Science*, *24*(7), 1089-1097.

Jiang, Q., Zhang, Y., & Pian, W. (2022). Chatbot as an emergency exist: Mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. *Information Processing and Management*, *59*(6), Article 103074.

Joosse, P. (2017). Leaderless resistance and the loneliness of lone wolves: Exploring the rhetorical dynamics of lone actor violence. *Terrorism and Political Violence*, *29*(1), 52-78.

Kiron, D., & Unruh, G. (2019). Even if AI can cure loneliness-should it? *MIT Sloan Management Review*, *60*(2), 1-4.

Krakovsky, M. (2018). Artificial (emotional) intelligence. *Communications of the ACM*, *61*(4), 18-19.

Kryshtaleva, M. K. (2017). The processes of alienation in the modern world and their features in visual culture. *AI and Society*, *32*(1), 117-120.

Langman, P. (2009). Rampage school shooters: A typology. *Aggression and Violent Behavior*, *14*(1), 79-86.

Lara, E., Martín-María, N., De la Torre-Luque, A., Koyanagi, A., Vancampfort, D., Izquierdo, A., & Miret, M. (2019). Does loneliness contribute to mild cognitive impairment and dementia? A systematic review and meta-analysis of longitudinal studies. *Ageing Research Reviews*, *52*, 7-16.

Larson, R., & Richards, M. H. (1991). Daily companionship in late childhood and early adolescence: Changing developmental contexts. *Child Development*, *62*(2), 284-300.

Lee, B., Kwon, O., Lee, I., & Kim, J. (2017). Companionship with smart home devices: The impact of social connectedness and interaction types on perceived social support and companionship in smart homes. *Computers in Human Behavior*, *75*, 922-934.

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, *122*, 875-888.

Luka Inc. (2023). *Creating a safe replika experience*. Replika. https://blog.replika.com/posts/creating-a-safe-replika-experience

Margolis, J. D., & Walsh, J. P. (2003). Misery loves companies: Rethinking social initiatives by business. *Administrative Science Quarterly*, *48*(2), 268-305.

Mazmanian, M., Orlikowski, W. J., & Yates, J. (2013). The autonomy paradox: The implications of mobile email devices for knowledge professionals. *Organization Science*, *24*(5), 1337-1357.

Merrill Jr, K., Kim, J., & Collins, C. (2022). AI companions for lonely individuals and the role of social presence. *Communication Research Reports*, *39*(2), 93-103.

Metz, C. (2020). Riding out quarantine with a chatbot friend:'I feel very connected.'. *The New York Times*.

Metz, C., & Schmitz, G. (2023). *Elon Musk and Others Call for Pause on A.I., Citing "Profound Risks to Society."* The

New York Times. https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html

Monteiro, E., Constantinides, P., Scott, S., Shaikh, M., & Burton-Jones, A. (2022). Qualitative research methods in information systems: a call for phenomenon-focused problematization. *MIS Quarterly*, *46*(4), i-xviii.

Montemayor, C., Halpern, J., & Fairweather, A. (2022). In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *AI and Society*, *37*(4), 1353-1359.

O'Reilly, C. A., & Tushman, M. L. (2013). Organizational ambidexterity: Past, present, and future. *Academy of Management Perspectives*, *27*(4), 324-338.

Palgi, Y., Shrira, A., Ring, L., Bodner, E., Avidor, S., Bergman, Y., Cohen-Fridel, S., Keisari, S., & Hoffman, Y. (2020). The loneliness pandemic: Loneliness and other concomitants of depression, anxiety and their comorbidity during the COVID-19 outbreak. *Journal of Affective Disorders*, *275*, 109-111.

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, Article 107600.

Pentland, B. T. (1999). Building process theory with narrative: From description to explanation. *Academy of Management Review*, *24*(4), 711-724.

Picard, R. W. (1995). Affective Computing. *M.I.T. Media Laboratory Perceptual Computing Section Technical Report*(321).

Porter, A. J., & van den Hooff, B. (2020). The complementarity of autonomy and control in mobile work. *European Journal of Information Systems*, *29*(2), 172-189.

Possati, L. M. (2022). Psychoanalyzing artificial intelligence: the case of Replika. *AI and Society*.

Purtill, J. (2023). *Replika users fell in love with their AI chatbot companions. Then they lost them.* ABC News. https://www.abc.net.au/news/science/2023-03-01/replika-users-fell-in-love-with-their-ai-chatbot-companion/102028196

Quartz. (2017). *Replika: This app is trying to replicate you.* Quartz: Machines With Brains. https://qz.com/1698337/replika-this-app-is-trying-to-replicate-you

Riemer, K., & Peter, S. (2021). Algorithmic audiencing: Why we need to rethink free speech on social media. *Journal of Information Technology*, *36*(4), 409-426.

Rook, K. S. (1987). Social support versus companionship: effects on life stress, loneliness, and evaluations by others. *Journal of Personality and Social Psychology*, *52*(6), 1132.

Rowe, F., Ngwenyama, O., & Richet, J. L. (2020). Contact-tracing apps and alienation in the age of COVID-19. *European Journal of Information Systems*, *29*(5), 545-562.

Sahu, O. P., & Karmakar, M. (2022). Disposable culture, posthuman affect, and artificial human in Kazuo Ishiguro's Klara and the Sun (2021). *AI & Society*, 1-9.

Seymour, M., Yuan, L. I., Dennis, A., & Riemer, K. (2021). Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, *22*(3), 9.

Sharkey, A., & Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics and Information Technology*, *23*(3), 309-316.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human–chatbot relationships. *International Journal of Human Computer Studies*, *168*, Article 102903.

Somers, M. (2019). *Emotion AI, explained*. MIT Sloan Management Review. https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained

Sorkin, D., Rook, K. S., & Lu, J. L. (2002). Loneliness, lack of emotional support, lack of companionship, and the likelihood of having a heart condition in an elderly sample. *Annals of Behavioral Medicine*, *24*(4), 290-298.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research*, *22*(3), Article e16235.

Tilson, D., Lyytinen, K., & Sørensen, C. (2010). Research commentary—Digital infrastructures: The missing IS research agenda. *Information Systems Research*, *21*(4), 748-759.

Troya, M. I., Babatunde, O., Polidano, K., Bartlam, B., McCloskey, E., Dikomitis, L., & Chew-Graham, C. A. (2019). Self-harm in older adults: systematic review. *The British Journal of Psychiatry*, *214*(4), 186-200.

Upper Echelon. (2022). *Replika - A Mental Health Parasite*. https://www.youtube.com/watch?v=hUQNiy4K7VU

Vaast, E. (2023). Strangers in the dark: Navigating opacity and transparency in open online career-related knowledge sharing. *Organization Studies*, *44*(1), 29-52.

Vogel, D. (2007). *The market for virtue: The potential and limits of corporate social responsibility*. Brookings Institution Press.

Wareham, J., Fox, P. B., & Giner, J. L. C. (2014). Technology ecosystem governance. *Organization Science*, *25*(4), 1195-1215.

Xie, T., & Pentina, I. (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, Hawaii, USA.

Yao, T., Zheng, Q., & Fan, X. (2015). The impact of online social support on patients' quality of life and the moderating role of social exclusion. *Journal of Service Research*, *18*(3), 369-383.