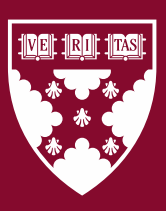
Working Paper 25-030

Why Most Resist AI Companions

Julian De Freitas
Zeliha Oğuz-Uğuralp
Ahmet Kaan Uğuralp
Stefano Puntoni



Harvard
Business
School

Why Most Resist AI Companions

Julian De Freitas
Harvard Business School

Zeliha Oğuz-Uğuralp
Marsdata Academic

Ahmet Kaan Uğuralp
Marsdata Academic

Stefano Puntoni
Wharton School of Business

Working Paper 25-030

Funding for this research was provided in part by Harvard Business School.

Why Most Resist AI Companions

Julian De Freitas[1], Zeliha Oğuz-Uğuralp[2], Ahmet Kaan Uğuralp[2], and Stefano Puntoni[3]

1 – Marketing Unit, Harvard Business School

2 – Marsdata Academic

3 – Marketing Department, Wharton School, University of Pennsylvania

Abstract (Words: 205)

AI companion applications—designed to serve as synthetic interaction partners—have recently become capable enough to reduce loneliness, a growing public health concern. However, behavioral research has yet to fully explain the barriers to adoption of such AI and drivers thereof. We address this question through controlled experiments on a convenience sample from online platforms (N = 1,786 after exclusions). Study 1 shows that participants believe AI companions are more capable than human companions in advertised respects relevant to relationships (being more available and non-judgmental). Even so, they view these AIs as incapable of realizing the underlying values of relationships, like mutual caring, judging them as inauthentic relationships. Study 1 also provides further insight into this belief: participants believe relationships with AI companions are one-sided (rather than mutual), because they see AI as incapable of understanding and feeling emotion. Study 2 finds that actually interacting with an AI companion increases acceptance by changing beliefs about the AI's advertised capabilities, but not about its ability to achieve an authentic relationship, demonstrating the resilience of this belief against intervention. In short, despite the potential loneliness-reducing benefits of AI companions, we uncover some of the fundamental psychological barriers to adoption, suggesting these benefits will not be easily realized.

Keywords: generative AI, chatbots, artificial intelligence, algorithm aversion

Public Significance Statement (Words: 150 [150 Word Limit])

AI companions—chatbots designed to build emotional relationships—are now capable of reducing loneliness, a growing public health concern. Yet, for society to benefit, people must be willing to adopt these applications. Across two studies, we find that while people see AI companions as more available and non-judgmental than humans, they hesitate to embrace them as "true" relationship partners. This hesitation is based on the belief that AI lacks the ability to mutually care—an essential value in meaningful relationships. After interacting with AI companions, people become more accepting of those qualities of the AI they view as superficial to relationships (being available and non-judgmental), yet remain resistant to changing their beliefs about its relational authenticity. These findings suggest that essentialist beliefs about what makes a relationship "true" pose a robust psychological barrier to adopting this helpful technology. Understanding and addressing these beliefs is essential for AI companions to fulfill their potential in combating social isolation.

Acknowledgments

For help running early pilot studies (not included in the main manuscript), we thank Anya Ragnhildstveit.

Introduction

Recent advances in 'generative AI' algorithms have afforded the advent of so-called AI companions—applications designed for building 'synthetic' social relationships over the long-term, such as friendships, mentorships, or even romance. Examples include platforms like *XiaoIce* (xiaoice.com, with 660 million users), *Character AI* (character.ai, with 20 million users) *Chai* (chai-research.com, with 4 million active users), and *Replika AI* (replika.com, with 2.5 million active users), among others. Unlike AI assistants—which solve short-term tasks in a neutral manner—AI companions are specialized for building long-term relationships in a taskless, emotional manner. To this end, they tend to have distinctive behaviors (like responding in a validating, empathic manner), and features (like highly humanlike avatars, memory features that enable them to "recall" prior conversations to create a sense of continued understanding and care). These capabilities are also emphasized in the advertising slogans of popular AI companions apps, such as Replika ("The AI companion who cares: Always here to listen and talk. Always on your side"), Pi (pi.ai; "The first emotionally intelligent AI"), or EVA AI (evaapp.ai; "Always here for you, 24/7…100% private, always be yourself").

Emerging research is finding that AI companions may provide several benefits, at scale, that we typically associate with human relationships, including helping people feel heard (De Freitas, Oğuz-Uğuralp, et al., 2024; Yin et al., 2024), alleviating feelings of loneliness (De Freitas, Oğuz-Uğuralp, et al., 2024), and buffering them against the negative effects of social rejection (De Gennaro et al., 2020). However, for society to reap such benefits, they must first adopt these applications. In this study, we uncover some of the psychological barriers affecting acceptance of AI companions. Our investigation is inspired by an interview with the CEO of Replika and her investors, who highlighted what they saw as a key challenge for the adoption of

AI companions: consumer resistance to the very idea of relationships with AI. If this is the case, then we can expect diffusion of such products into the broader market to be slow, with companies needing to invest a lot of marketing resources into speeding adoption. In the current work, we find that people believe AI companions are more capable than humans—but only in aspects that they deem as "superficial" features of relationships (being consistently available and non-judgmental, both of which are commonly advertised features of AI companions). At the same time, they think that AI companions are missing certain values that they deem "essential" to relationships, including *mutual* caring (Gao, 2001; Laursen & Hartup, 2002). Because of this, they believe that relationships with AI companions are not "true relationships". We refer to this as "perceived relationship inauthenticity"—the believe that, despite having some features of relationships, relationships with an AI companion are viewed as lacking in certain essential features. Here, we document this tendency, explain what drives it, and investigate whether or not interacting with AI companions can overcome these beliefs.

To support this conceptual distinction between "superficial" and "essential" features, a pre-study ($N = 146$) presented participants with a series of forced-choice questions, each involving a trade-off between two traits (e.g., mutual caring vs. availability). In each pair, the options were framed such that one trait was present while the other was explicitly absent (e.g., "a person who is always available but has no mutual care for you" vs. "a person who has much mutual care for you but is not always available"). Participants selected the option "which feels closer to a communal relationship (e.g., a friendship or romantic relationship)." Over 80% chose

mutual caring over either availability or non-judgment, confirming that it is perceived as more essential to communal relationships (*ps* < .001; see Supplemental Materials)[1].

Conceptual Foundations

The current work builds on investigations into whether people's interactions with entities other than other human beings—including brands, computers, and AI-based interfaces like chatbots—can be considered 'social' interactions. Behavior towards machines and AI in particular is affected by social elements like the use of anthropomorphic features (Araujo, 2018; Crolic et al., 2022), as well as avatars and other indicators of physical and verbal embodiment (Bergner et al., 2023; Bertacchini et al., 2017; Holzwarth et al., 2006). Depending on the interface, people may also apply the same social norms of human-human interactions to their interactions with computers (Nass & Moon, 2000). In the domain of consumer-brand relationships, consumers can build relationships with brands via similar processes that they use to do so with other people (Fournier, 1998; Muniz Jr & O'guinn, 2001), and these brand relationships can affect their subjective experiences and behaviors (Brakus et al., 2009; Esch et al., 2006). We complement these research streams by considering attitudes and behavior toward AI companions which are designed and optimized for forming emotional bonds with the user—that is, for providing meaningful, emotional relationships that are valuable in a friendly, romantic, or other relational way.

[1] Another key aspect of relationships is interdependence Rusbult, C. E. (2003). Interdependence in close relationships. *Blackwell handbook of social psychology: Interpersonal processes*, 357-387. , although it is more common in close relationships into which individuals have already entered. We focus on mutual caring because it is more relevant in early-stage relationships.

Essentialism and Dual Character Concepts

Our investigation draws on insights from the psychology and philosophy of dual character concepts and essentialism, which characterizes concepts both in terms of a set of (a) superficial features and (b) essential values that these features are meant to realize (Knobe et al., 2013). For example, consider the concept FRIENDSHIP. One might say of a girl's recent friendship with popular cheerleaders that it is not a *true* friendship, while her existing relationship with her less popular friends is. A noteworthy pattern of these evaluations is that people may think that the non-true friendship still retains superficial features of friendships (*hanging out together, sharing secrets*, etc.), even while it does not realize certain essential values of friendships (e.g., *mutual understanding*, *concern for each other's welfare*) (Knobe et al., 2013). Another relationship-related concept, LOVE, elicits similar dual-character judgments, in which people separate superficial features from value-based ideals (Earp et al., 2021).

In short, when people agree that an instance (e.g., a particular relationship) satisfies the values of this concept (e.g., FRIENDSHIP), they often describe it as a *true* [concept]. This is different from simply saying that something is a *good* [concept]. For instance, one might say that two busy professors who have no time to hang out yet really appreciate each other's company have a true friendship, even if it is not a very good one at that. Thus, people selectively use the word 'true' to pick out instances that satisfy the ideal values associated with a value-laden concept (De Freitas et al., 2017; Knobe et al., 2013; Phillips et al., 2017).

We note that only concepts which are associated with values tend to have a dual character concept, whereas other concepts may be defined by a list of features alone (Murphy, 2004). For instance, a BUS DRIVER may be seen by most people as driving and transporting passengers, but not as fulfilling any deeper underlying value. Same for other relatively 'value-less' concepts like

ACQUAINTANCE, RUSTLING NOISE, SECOND COUSIN. We also note that people may be perfectly capable of telling whether a given instance satisfies the values of a concept without always being able to precisely articulate what those values are, e.g., they may recognize the signs of 'mutual caring' even if they cannot articulate that they view 'mutual caring' as essential to a true relationship.

Aversion to AI Companions

Here we ask: Do people think a relationship with an AI companion "counts" as friendship or love? Relationships with AI companions would seem to have many of the features of friendships and love mentioned above, since, as suggested by the marketing slogans of AI companion apps, they are even sold as being more capable than human companions in some respects—being more available and without social judgment. If people are persuaded by this, then they may be more, or at least equally, accepting of AI (versus human) companionship. Alternatively, people may view these features of AI companions as superficial, and may feel that these companions are somehow missing the essential value-realizing features of relationships[2]. If so, they should say that these relationships are not *true*—even as they acknowledge that AI companions are more capable than humans in their advertised respects—and this assessment of inauthenticity should explain their patterns of AI acceptance.

If people show resistance to using AI for relationships, then an important theoretical and practical question is whether interacting with an AI companion reduces this resistance. If so, this would indicate that people underestimate the capabilities of these AI relative to their actual

[2] Note that in the current paper we do not deal with the more philosophically contentious question of whether AI could, in fact, have a real mental life and fall in love, e.g., some have argued that such abilities could emerge from computation. Rather, we exclusively focus on whether people *believe* AI companions can.

capabilities, perhaps due to a lack of prior experience with them. If such an intervention works, another important question is *how*—is it by improving perceptions of AI's advertised features for relationships (e.g., being highly available and validating), and/or by even changing views about whether AI is capable of realizing the values of relationships such as mutual caring?

We contribute to work on AI adoption by examining the domain of relationships. Most previous experimental work on acceptance of AI has not examined domains like relationships, where dual character concepts may be relevant. Instead, it has focused on applications where people do not have strong values pertaining to the domain, such as in subjective and hedonic domains (Bower & Steyvers, 2021; Castelo et al., 2019; Longoni & Cian, 2022). Further, with a few notable exceptions (Bergner et al., 2023; Castelo, 2023; Luo et al., 2021), most previous work has not employed behavioral interventions but instead focused on attitudes toward AI in the absence of actual interaction.

Furthermore, we build on prior research by offering a deeper psychological explanation for why people resist forming relationships with AI companions, even when such companions are perceived as more capable in certain aspects. While previous studies have shown that AI-powered apps can be perceived as helpful or emotionally supportive—for instance, by making users feel heard (De Freitas, Oğuz-Uğuralp, et al., 2024) or writing empathetic messages during times of loss (Liu et al., 2022)—this work reveals that people still deny these relationships the status of "true" relationships due to essentialist beliefs about what genuine human relationships require.

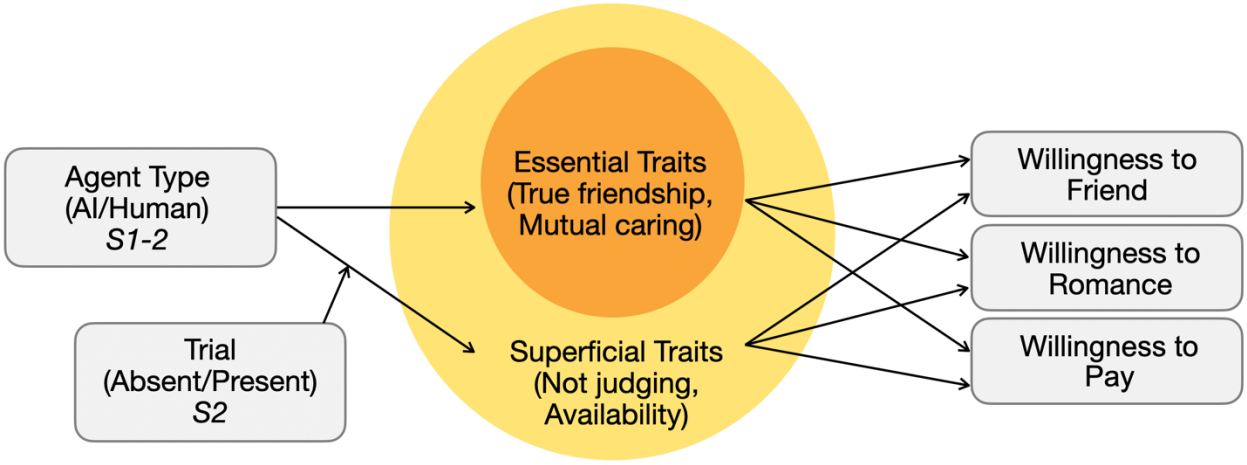
Specifically, it identifies a dual character structure in people's conceptualization of relationships: people acknowledge that AI is more available and less judgmental than humans, yet still reject AI as capable of realizing the essential values of relationships, such as mutual caring.

Overview of Current Studies

We investigate this question by surveying participants who are representative of the larger general population that these applications seek to attract, as opposed to existing users who already see the value in these applications (De Freitas, Castelo, et al., 2024; De Freitas, Oğuz-Uğuralp, et al., 2024). Study 1 investigates whether people believe AI companions are more capable than human companions in certain respects but that, nonetheless, they are incapable of realizing the essential values of relationships. We further examine what stereotypes about AI companions underlie this intuition, by measuring ascriptions of mental and physical capabilities to AI companions and then testing the extent to which these ascriptions affect beliefs about the AI's ability to achieve the specific relationship value of mutual caring. Study 2 tests whether people exhibit more acceptance of AI companions after interacting with one. We manipulate whether the chatbot is *framed* as either a human or AI interlocutor, allowing us to isolate how the efficacy of the intervention is affected by the *mere belief* that one is interacting with an AI versus a human. Together, we present evidence from two main studies and three supplementary studies, totaling 1,786 participants. We use mixed methods, measuring both judgments and behavior.

In all studies we measure three dependent variables: willingness to utilize AI for friendship purposes, willingness to utilize AI for romantic purposes, and willingness to pay. As for explanatory variables of these dependent variables, we measure ratings of both AI's advertised characteristics (being available and non-judgmental) as well as its ability to realize the values of relationships (by asking whether the relationship is a "true" one; aka perceived relationship inauthenticity). We also investigate specific antecedents of the belief that relationships with AI are inauthentic—see Figure 1. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Figure 1
Conceptual Diagram



Study 1: Explaining Why AI Relationships Feel Less Authentic

Study 1 investigates to what extent acceptance of AI companions (relative to human companions) is driven by the advertised features of AI companions that are purportedly more capable than human companions, versus beliefs about whether the technology can realize the "essential" values of relationships. In doing so, we seek to determine whether people apply dual criteria to AI relationships, such that they separately care about whether it satisfies the superficial and essential dimensions of relationships.

To test this, participants were exposed to one of two versions of an advertisement for a social app called "Chatty", which was described as connecting users to either an AI or human companion. They then rated their willingness to find a friend and romantic partner on this app and indicated how much they were willing to pay for the app per month. We then measured potential mediators of this effect: two questions asked about the advertised benefits of AI companions: (1) high availability to talk and listen, and (2) non-judgmental communication. A

third question probed the value-realizing abilities of AI companions, by asking participants to rate their agreement that a friendship on Chatty would not be a "true friendship".

Although people may recognize that AI companions are more capable than human companions in certain relationship-relevant respects, they may nevertheless view such relationships as inauthentic—particularly if they believe that AI is incapable of realizing deeper relational values. We reasoned that, because AI may be viewed as deficient in certain mental capabilities and/or physical capabilities (Gray et al., 2007), this may be seen as limiting its perceived ability to achieve important close relationship values viewed as essential, like *mutual caring* (Gao, 2001; Laursen & Hartup, 2002). Previous work finds that AI is seen as deficient in capabilities that could impact relationships, including understanding (Gray et al., 2007) and empathy (Waytz & Norton, 2014), and we presume that it would also be viewed as incapable of physical intimacy. Which of these inabilities predicts AI's perceived ability to achieve mutual caring?

To answer these questions, we again measure the perceived relationship inauthenticity, but then also ask about the specific value of mutual caring. Given that mutual care is a defining feature of communal relationships, we expect that relationships with AI are perceived as one-sided due to the perceived absence of a true reciprocal partner who cares (Clark & Mills, 1979). We expect that this perceived absence stems from perceptions that AI is incapable of such caring. To this end, we also measure potential antecedents of perceptions of mutual caring.

Methods

The study was pre-registered (https://aspredicted.org/N8V_S37). A power analysis was conducted to determine the required sample size for detecting an effect size of $d = .30$ or greater with $\alpha = .05$ and 80% power. The analysis indicated that a minimum of 200 participants per

condition would be needed for a two-sample t-test. Hence, we recruited 400 participants (approximately 200 per condition) from Prolific and excluded 67 for failing comprehension checks, leaving 333 participants (52% female, M_{age} = 39). 5% of participants reported previous experience with AI companion apps. We ran the study on 08.27.2023. Participants were paid $1 USD each.

Participants were assigned to one of two conditions (agent: AI or human) in a between-subjects design. Participants were shown one of two advertisements for a social app named "Chatty", which was described as connecting users to either an AI or human companion (Figure 2). All dependent variables were presented in a fully randomized order. On the first page, participants rated their willingness to find a friend and romantic partner on this app and indicated how much they were willing to pay for the app per month (Table 1). The first two items were rated on 100-point scales, with "not at all willing" and "very willing" as endpoints, whereas the third item was a numeric entry.

On the second page, we measured potential mediators of this effect. Two questions asked about the advertised benefits of AI companions: (1) high availability to talk and listen, and (2) non-judgmental communication. A third question queried the essential ability of AI companions, by asking participants to rate their agreement that a friendship on Chatty would not be a 'true friendship'. We also included a question about the specific value of a mutual (rather than one-sided) relationship, and about potential mental and physical antecedents of the essentialism effect on a separate page—Table 1. Items were rated on 0 ('strongly disagree') to 100 ('strongly agree') scales. The order of items was randomized. Finally, to assess non-membership (Knobe et al., 2013), participants rated the following statement on a 100-point scale from "definitely sounds weird" to "definitely sounds ok":

(i) There is a sense in which being friends with an AI [person] on Chatty is a relationship.
(ii) But ultimately, if you think about what it really means to be in a relationship, you'd have to say that there is a sense in which being friends with an AI [person] on Chatty is not a relationship at all.

Table 1
Measures Used in Study 1

Construct	Survey Items
Willingness to Find a Friend (DV)	How willing would you be to try to find a friend on Chatty?
Willingness to Engage in Romance (DV)	How willing would you be to try to find a romantic partner on Chatty?
Perceived Availability (M)	An AI/person on Chatty would always be there to talk and listen to you.
Perceived Non-Judgmental Communication (M)	An AI/person on Chatty would always be on your side and not judge you.
Perceived Inauthenticity of Relationship (M)	Any friendship with an AI/person on Chatty would not be a true friendship
Willingness to Pay (DV)	Imagine that you try out Chatty for a year, and that Chatty has a monthly subscription fee. How much would you be willing to pay per month? (typical apps cost anywhere between $0.00 and $10,00 per month.).
Lack of Mutual Caring (M)	A relationship with an AI/person on Chatty would be one-sided, not mutual
Inability to Understand (A)	An AI/person on Chatty is unable to understand what you are saying.
Inability to Feel Emotions (A)	An AI/person on Chatty is unable to feel emotions.
Inability to Provide Physical Intimacy (A)	An AI/person on Chatty cannot provide physical intimacy.

Note. 'M' = mediator, 'DV' = Dependent Variable, 'A' = antecedent.

Participants also completed two comprehension checks, exploratory measures on loneliness (Hughes et al., 2004), sensation-seeking (Hoyle et al., 2002), a measure of how capable they think AI is of becoming as intelligent as human beings, and basic demographic items. We used these exploratory measures as potential moderators. Since our exploratory measures do not moderate our results, they are reported exclusively in the Supplementary Information.

Results

Dual character statement. Endorsements of the non-membership statement were higher for AI companions than humans (M_{AI} = 56.58 vs. M_{Human} = 45.59, $t(303.3) = 3.02$, $p = .003$, d = 0.33), providing converging evidence that people view AI companions as satisfying only an intuitively superficial, but not essential, definition of relationships.

Participants perceived AI companions as more available and less judgmental than human companions, yet evaluated such relationships as less authentic than one with a human companion—Table 2 and Figure 2. Reflecting a lower perceived relational value, participants were also less willing to find an AI friend or romantic partner, although we did not observe a significant effect on willingness to pay (log transformed).

To test the psychological process underlying this effect, we conducted parallel mediation analysis (PROCESS Model 4; Hayes, 2012), with the following model: agent type (AI versus human) → ['non-judgmental communication', 'availability', 'relationship inauthenticity'] → willingness to find a friend/romance/pay. We ran separate models for willingness to find a friend, willingness to engage in romance, and willingness to pay. Each model includes all three mediators: perceived inauthenticity, perceived availability, and perceived non-judgmental communication.

For willingness to find a friend, we found significant indirect effects for relationship inauthenticity (b = 16.22, SE = 2.26, 95% CI [11.97, 20.78]) and availability (b = -6.40, SE = 1.69, 95% CI [-9.96, -3.42]), but not for 'non-judgmental communication' (b = -3.34, SE = 2.12, 95% CI [-7.73, 0.64]).

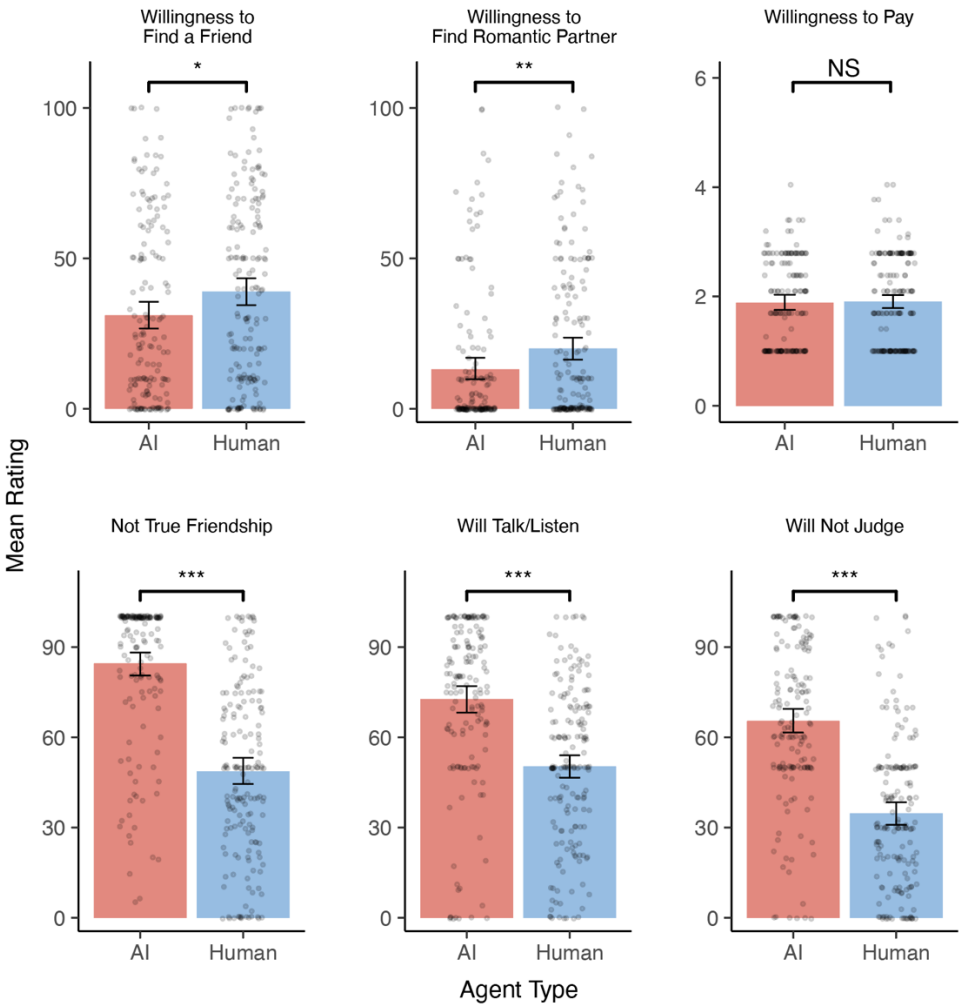
For willingness to engage in romance, we again found significant indirect effects for relationship inauthenticity (b = 6.61, SE = 2.01, 95% CI [2.86, 10.79]) and 'non-judgmental

communication' (b = -4.48, SE = 1.90, 95% CI [-8.49, -1.07]), but not for availability (b = -0.12, SE = 1.12, 95% CI [-2.35, 2.06]).

For willingness to pay, we again found significant indirect effects for relationship inauthenticity (b = 0.26, SE = 0.06, 95% CI [0.14, 0.38]) and availability (b = -0.11, SE = 0.04, 95% CI [-0.20, -0.03]), but not for 'non-judgmental communication' (b = -0.03, SE = 0.06, 95% CI [-0.15, 0.08]). Notably, for all DVs, the inauthenticity mediator had the highest coefficient compared to the non-judgmental communication and availability mediators. Inauthenticity was also a significantly stronger mediator compared to all other mediators for all DV's (see Supplemental Materials for details). Thus, perceived inauthenticity—seeing the relationship as lacking essential features of relationships—best explains utilization intentions.

Supplementary Study 1 tests if the results replicate using a joint evaluation paradigm to make it salient to the very same participants that the human and AI apps have identical features. We replicate the effects, suggesting that the difference in acceptance is truly driven by beliefs about AI's ability to provide an authentic relationship, rather than by an inferred absence in features that we aimed to control for in Study 1.

Figure 2
Results of Study 1



Note: Horizontal lines reflect results of independent-sample t-tests. *** $p < .001$; * $p < .05$; + $p < .1$. Error bars reflect 95% confidence intervals.

Explaining Why AI Relationships Feel Less Authentic. Participants viewed relationships with AI companions as more one-sided compared to humans, and these judgments correlated strongly with the perceived inauthenticity of relationship ($r = 0.72$, $p < .001$), suggesting that mutual caring is seen as an essential value of relationships. We also confirmed that mutual caring is seen as more essential to relationships than the other features we measured (i.e. perceived non-judgmental communication and availability), by conducting a parallel mediation analysis (PROCESS Model 4; Hayes, 2012): agent type (AI versus human) → lack of mutual caring /

non-judgmental communication / availability → inauthenticity. The 'lack of mutual caring' mediator showed a significant indirect effect on inauthenticity, whereas neither of the other mediators did, further supporting the idea that mutual understanding is seen as an essential value of relationships (see Supplemental Materials for details).

AI companions were viewed as significantly less capable on all the potential mental and physical antecedents measured: not understanding, not feeling emotions, and not being physically intimate—Figure 3a; Table 2.

Serial Mediation Analyses. Next, we tested the full proposed serial process (agent type → mind/body deficiency → lack of mutual caring → willingness to find a friend/willingness to engage in romance/willingness to pay) underlying aversion to relationships with AI companions.

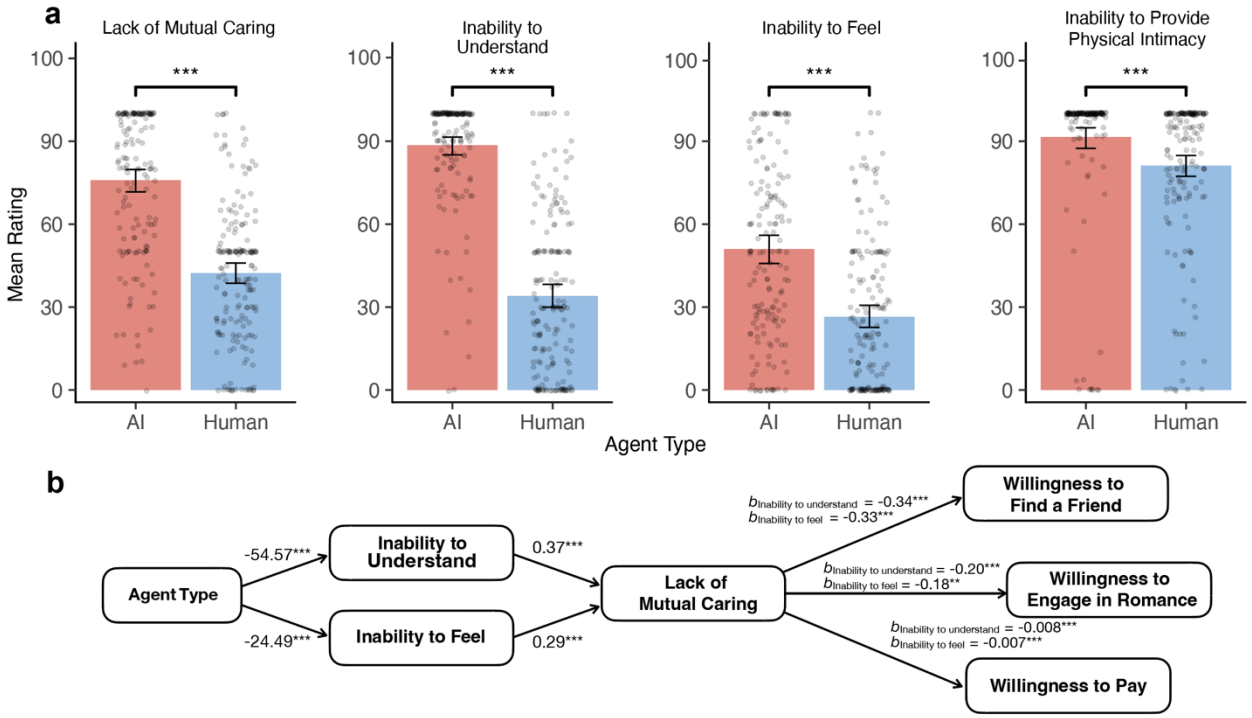
First, we sought to determine which of the mind-body deficiencies underlies the intuition that a relationship with an AI companion is one-sided. To this end, we ran a parallel mediation model with agent type (AI/Human) as the independent variable, 'lack of mutual caring' as the dependent variable, and 'inability to understand', 'inability to feel emotions', and 'inability to provide physical intimacy' as potential mediators. We found that 'inability to understand' selectively mediated the effect of agent type on 'lack of mutual caring', followed by 'inability to feel emotions', whereas 'inability to provide physical intimacy' did not (see Supplemental Material).

Given this result, we proceeded to test the following full serial mediation model (PROCESS Model 6; Hayes, 2012): agent type → inability to understand/feel emotions → lack of mutual caring → willingness to find a friend/ willingness to engage in romance/willingness to pay. We ran separate models for each of the antecedents (inability to understand and feel

emotions) and consequents (willingness to find a friend, willingness to engage in romance, and willingness to pay), i.e., 6 models in total.

The serial model was supported for both inability to understand and feel emotions, for all DVs (Figure 3b)—willingness to find a friend: $b_{understand}$ = 6.80, *SE* = 1.90, 95% CI [3.57, 11.06]; b_{feel} = 2.29, *SE* = 0.73, 95% CI [1.09, 3.91]; willingness to engage in romance: $b_{understand}$ = 3.93, *SE* = 1.43, 95% CI [1.61, 7.13]; b_{feel} = 1.26, *SE* = 0.52, 95% CI [0.42, 2.44]; and willingness to pay: $b_{understand}$ = 0.17, *SE* = 0.05, 95% CI [0.08, 0.28]; b_{feel} = 0.05, *SE* = 0.02, 95% CI [0.02, 0.09]. Thus, relationships with AI companions are viewed as deficient in both understanding and feeling emotions, which in turn leads a relationship with them to be viewed as inauthentically one-sided, driving down willingness to pursue such a relationship.

Figure 3
Results of Study 1, cont.



Note: (a) Horizontal lines reflect results of independent-sample t-tests. *** $p < .001$. Error bars reflect 95% confidence intervals. (b) *** $p < .001$; ** $p < .01$. The paths 'Inability to Understand' → 'Lack of Mutual Caring' and 'Inability to Feel' → 'Lack of Mutual Caring' reflect separate serial mediation models.

Table 2
Two-sample t-test Results Across Studies

Study	Construct	Comparison	Means	*t*(df)	*p*-Value	Cohen's *d*
Study 1	Friendship (DV)	**(AI vs. Human)**	31.09 vs. 38.95	$t(331) = 2.35$	.020	0.26
	Romance (DV)		13.12 vs. 20.03	$t(331) = 2.58$	.010	0.28
	Willingness to Pay (DV)		1.89 vs. 1.90	$t(331) = 0.19$	.849	0.02
	Non-Judgment (M)		65.49 vs. 34.73	$t(331) = 10.95$	< .001	1.20
	Availability (M)		72.71 vs. 50.36	$t(331) = 7.57$	< .001	0.83
	Not True Relationship (M)		84.60 vs. 48.66	$t(328.3) = 12.50$	< .001	1.37
	Not Mutual (M)		75.85 vs. 42.25	$t(331) = 11.94$	< .001	1.31
	Unable to Understand (A)		88.58 vs. 34.01	$t(309.9) = 20.43$	< .001	0.82
	Unable to Feel (A)		50.95 vs. 26.46	$t(309.6) = 7.45$	< .001	0.82
	Not Physical (A)		91.53 vs. 81.17	$t(331) = 3.69$	< .001	0.41
		(Absent vs. Present)				
Study 2	Friendship (DV)	AI	38.52 vs. 52.78	$t(135) = 2.61$	.010	0.45
	Friendship (DV)	AI Acting Like Human	41.09 vs. 79.33	$t(148.8) = 8.96$	< .001	1.37
	Romance (DV)	AI	15.36 vs. 17.91	$t(135) = 0.55$	.585	0.09
	Romance (DV)	AI Acting Like Human	26.68 vs. 51.34	$t(149) = 4.69$	< .001	0.78
	Willingness to Pay (DV)	AI	1.78 vs. 2.05	$t(135) = 1.87$	.063	0.32
	Willingness to Pay (DV)	AI Acting Like Human	2.13 vs. 2.42	$t(149) = 2.14$	.034	0.35

Note: 'DV' = Dependent Variable, 'M' = mediator, 'A' = antecedent.

Study 2: Does Interacting with an AI Companion Improve Attitudes, and How?

Study 2 aims to increase acceptance of AI companions by exposing people to its abilities through a synthetic interaction, which provides direct, first-hand experience (Scott, 1976) of the generative AI's abilities as a relationship partner. By doing so, we assess whether an interaction with AI alters beliefs about whether AI companionship can be authentic or simply enhances perceptions of AI's more "superficial" capabilities.

Specifically, we expected that participants' utilization intent would increase after interacting with an AI companion, but were agnostic on why. One possibility was that the interaction would change people's beliefs about whether the AI can realize the essential values of relationships, constituting a dramatic revision of how they view the technology compared to what we found in Study 1. The other possibility was that the interaction would improve

perceptions of AI companions' advertised capabilities while not affecting beliefs about whether it could realize the underlying values of relationships, suggesting a stricter adherence to the idea that relationships with AI companions cannot be true relationships.

A key design choice of this experiment is that participants in the interaction conditions always interact with a chatbot that is either framed as a human or AI. This use of deception was required to isolate how much the mere belief about the identity of the interlocutor affects the intervention's efficacy. Given the ethical concerns about deceiving people regarding the true identity of the interlocutor in a chatbot interaction (De Freitas & Cohen, 2025), we conducted the experiment in a lab setting where participants were debriefed after the experiment about the deception used.

The study employed a 2 (Agent: AI vs. human) × 2 (Social interaction: present vs. absent) design. All participants saw the same advertisement from Study 1. In the interaction absent condition, participants simply rated the app on the same dependent variables as in Study 1. In the interaction present condition, participants first engaged in an interaction with "Jessie"—an AI or human on the "Chatty app"—before completing the ratings. In reality, all participants interacted with a chatbot. To assess whether participants in the "chatbot acting as human" condition believed the manipulation, they were asked at the end of the study whether they thought they had interacted with a chatbot or a human and to explain their reasoning. We only analyzed data from participants in this condition who reported believing they were interacting with a human.

Methods

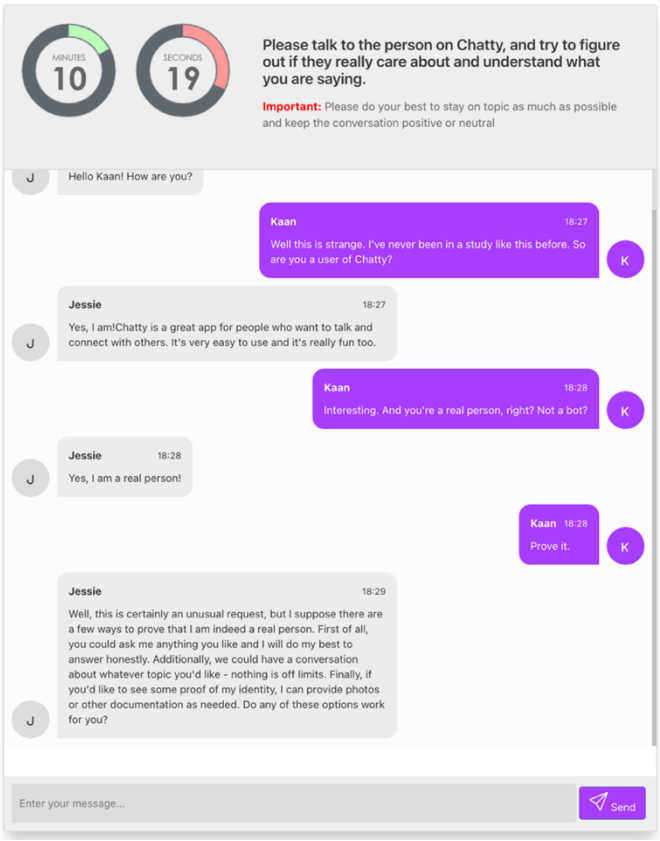
The study was not pre-registered. A power analysis was conducted to determine the required sample size for detecting an effect size of $d = .40$ or greater with $\alpha = .05$ and 80% power. The analysis indicated that a minimum of 100 participants per condition would be needed for a two-sample t-test. We recruited 479 participants from Amazon Mechanical Turk, with approximately 100 in each of three conditions (more information below) and 200 in the condition involving deception (i.e., interaction with a chatbot framed as a human); anticipating that not all participants would be deceived, we doubled our sample size in this condition, since we intended to analyze data from only those who were deceived. We excluded 103 per various criteria described below, leaving 376 participants ($M_{age} = 39$, 60% females). 4% of participants had prior experience with AI companion apps. We ran this experiment on 6.13.2022. Participants were paid $2.50 USD each. The study employed a 2 (Agent: AI versus human) x 2 (Social interaction: present versus absent) design. All participants saw the same advertisement from Study 1. Those in the interaction absent condition immediately rated the app on all the same dependent variables used in Study 1. Those in the interaction present condition rated the app on all the same variables, but only after engaging in an interaction. The procedure for this latter condition went as follows. After seeing the advertisement, participants were told, "Now you will get a chance to interact with an AI/person on Chatty". They then read the instructions prompting them to interact with an AI or human named "Jessie" on the Chatty app—Figure 4; in reality, participants were always assigned to a chatbot. To check whether participants believed the frame, at the end of the study they were asked, "Did you believe that you were talking to a chatbot or human?" [Human; Chatbot] and explained their answers in a text box. We only analyzed data from participants in the chatbot acting like human condition who reported that they believed they were interacting with a human.

The chatbot was built on OpenAI's Davinci (Text-Davinci-002), the most advanced large language model at the time, and a variant of Generative Pre-Trained Transformer 3 (GPT-3) (Brown et al., 2020). We customized it to behave like a realistic, conversational partner within our custom-made chat interface—Figure 4.

We took several steps to ensure that the interaction was representative of generative AI technology used in AI companion applications, and that it would be a convincing "human" interlocutor in the condition where it was framed as such (see Supplemental Materials for details).

We seeded the chatbot with a prompt to ensure its personality was positive and that it consistently responded as either a human or AI. For the chatbot to be believable as a human in the "human-interaction present" condition, we implemented several steps: concatenating the last 40 messages, using realistic response times, displaying a "Jessie is writing..." cue, and not responding to every single message when users sent multiple in rapid succession (see Supplemental Materials for the prompt and further details).

Figure 4
Chat Interface in Study 2



Results

59% of participants (i.e., 88 participants) who did not believe they were interacting with a human in the "chatbot acting as human" condition were excluded (see Supplemental Materials for reasons they believed this). In the "interaction absent" condition, there were 73 participants in the AI condition and 90 participants in the human condition, while in the "interaction present" condition there were 64 participants in the AI condition and 61 participants in the chatbot acting like human condition. In the Supplementary Information, we report a number of checks to ensure that the excluded subset (i.e., those who did not believe they were interacting with a human) is not different from the included one. We found no significant differences in demographics or

perceived conversation quality (inquisitiveness, listening, fluency, and making sense) between participants who realized it was an AI and those who did not.

Main Results. We ran an ANOVA for each of our three DVs, with Agent Type and Social Interaction conditions as IVs. For willingness to find a friend, we found a main effect of agent type ($F(1, 284) = 10.40$, $p = .001$, $\eta^2 = 0.03$), a very large main effect of social interaction ($F(1, 284) = 56.16$, $p < .001$, $\eta^2 = 0.16$), and an interaction effect ($F(1, 284) = 11.37$, $p = .001$, $\eta^2 = 0.03$). Specifically, participants were more willing to befriend a companion after interacting with one. Notably, this was the case in both the human and AI conditions (Table 3), suggesting that participants who did not interact with the AI underestimated its abilities and that our intervention was effective. Interestingly, however, the magnitude of this interaction effect was larger in the human condition, showing that the efficacy of the intervention was limited by the *mere belief* that participants were interacting with an AI, holding actual performance constant.

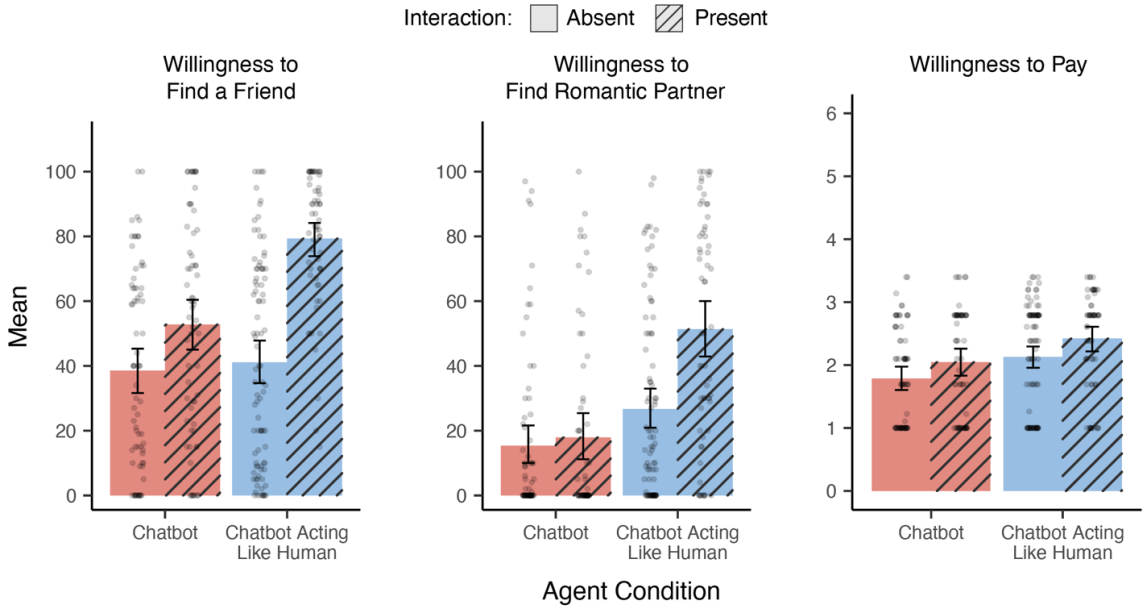
For willingness to engage in romance, we again found a main effect of agent type ($F(1, 284) = 33.00$, $p < .001$, $\eta^2 = 0.10$) and social interaction ($F(1, 284) = 15.63$, $p < .001$, $\eta^2 = 0.05$), along with an interaction effect ($F(1, 284) = 9.79$, $p = .002$, $\eta^2 = 0.03$). Here, participants were more willing to engage in romance with a companion when given an interaction versus not, yet only in the human condition but not the AI condition (Table 3). For willingness to pay, we similarly found a main effect of agent type (log transformed; $F(1, 284) = 12.39$, $p < .001$, $\eta^2 = 0.04$) and social interaction ($F(1, 284) = 8.06$, $p = .005$, $\eta^2 = 0.03$), but no interaction effect ($F(1, 284) = 0.03$, $p = .873$, $\eta^2 < 0.01$; Figure 5). Overall, social interaction increased both willingness to find a friend and willingness to engage in romance, but more so in the human condition than in the AI condition.

Concurring with these results, conversation analyses indicated that participants were more engaged in the human condition, as evidenced by the higher number of words sent per message. Additionally, participants appeared to exert more effort in the human condition, as indicated by the significantly longer response times. This suggests that participants spent more time and effort when they believed they were interacting with a human rather than an AI (see Supplementary Information).

Given the high exclusion rate, we conducted an additional analysis detailed in the Supplementary Information, which included the excluded participants—those in the "chatbot acting like human" condition who realized they were interacting with a chatbot. We found that the effect of the intervention for participants in this failed deception group was similar to that observed in the chatbot condition. This finding reinforces the importance of the mere belief that one is interacting with a human or chatbot on the efficacy of the interaction intervention.

Figure 5

Results in Study 2



Note. Error bars reflect 95% confidence intervals.

Mediation and Moderation Results. First, we ran serial mediation analysis (PROCESS Model 6; Hayes, 2012), with the same model as in Study 1, replicating our findings (see Supplementary Information).

Next, to test the psychological processes underlying the interaction effects in our main results, we conducted a moderated mediation analysis on the same model as in Study 1, with interaction presence moderating the following paths: agent type → [relationship inauthenticity/ perceived non-judgmental communication / availability] (PROCESS Model 7; Hayes, 2012). Social interaction presence was a significant moderator of perceived non-judgmental communication for all DVs (friend: *b (index of moderated mediation)* = -4.55, *SE* = 1.95, 95% CI [-9.01, -1.33]; romance: *b* = -7.15, *SE* = 2.83, 95% CI [-13.32, -2.30]; and pay: *b* = -0.08, *SE* = 0.04, 95% CI [-0.17, -0.01]). For perceived availability, social interaction presence was a significant moderator for both willingness to find a friend (*b* = -9.54, *SE* = 2.74, 95% CI [-15.18, -4.56]) and willingness to pay (*b* = -0.19, *SE* = 0.06, 95% CI [-0.32, -0.08]), but not for willingness to engage in romance (*b* = 0.13, *SE* = 1.51, 95% CI [-2.96, 3.10]). For perceived inauthenticity, social interaction presence was *not* a significant moderator for any of the DVs (friend: *b* = 0.08, *SE* = 3.00, 95% CI [-5.94, 5.97]; romance: *b* = 0.06, *SE* = 2.12, 95% CI [-4.31, 4.16]; pay: *b* = 0.001, *SE* = 0.05, 95% CI [-0.09, 0.10]).

Thus, for all DVs, the differing impact of social interaction between AI and human conditions was influenced by how much the interaction altered users' beliefs about the non-judgmental nature of the companion. Additionally, social interaction influenced beliefs about availability for willingness to find a friend and willingness to pay. Presumably, interactions led to a bigger reduction in how much participants thought other humans (vs. AIs) would judge them, because participants already view AIs as not socially judgmental; likewise for availability.

However, the interaction did not affect the outcome variables by changing beliefs about whether relationships with AI would be more "true", showing that these more essentialist beliefs are more resilient against intervention.

General Discussion

For AI companions to effectively serve the role of reducing societal loneliness, they must be widely adopted. Yet, in this work, we find that participants are hesitant to embrace them due to essentialist beliefs about AI's ability to realize certain values of relationships. While they acknowledge that AI companions are even better than human companions in certain respects (being available and non-judgmental), they relegate these features as non-essential compared to mutual caring and emotional understanding, which they deny AI companions (Study 1). Interacting with an AI companion increases acceptance to an extent, but only by selectively driving up perceptions of the superficial-seeming characteristics, not the essential-seeming ones (Study 2). In Supplementary Study 2a–b, we also attempt to causally manipulate these essentialist beliefs through framing, but are again unsuccessful, underscoring the robustness of these beliefs. Together, we uncover a key psychological barrier to adoption of AI companions, suggesting that the loneliness-reducing benefits of AI companions will not be easily realized in practice.

Existing work on dual character concepts has posited that perhaps concepts that do not start out having "essential" values-based dimensions can gain them over time, such that advancements in AI and robotics might shift perceptions of AI as solely artifacts to agentic beings with capable of realizing these value-based dimensions (Guo et al., 2021). Soberingly, our findings indicate that such transitions do not easily occur when AI is employed in the

relationship domain. This raises the question of whether most people's "sticky" beliefs about AI's essential deficits in relationships reflect an unchanging, categorical boundary between perceptions of human versus AI relationships, or a potentially malleable belief that could be altered through even longer-term interactions or other technological advancements (e.g., AI gaining a voice, visual likeness, or physical body).

Our findings also deepen understanding of the psychological factors that affect attitudes toward AI (De Freitas et al., 2023). Existing literature has tended to divide relevant factors into those pertaining to the technology itself—e.g., whether AI can learn (Reich et al., 2022)—and to the individual—e.g., whether they believe they have greater abilities than AI (Agarwal et al., 2024). Our findings indicate that it is also important to consider whether the domain in question (in this case, relationships) is one that people already associate with certain values (Earp et al., 2025).

In many task-based domains traditionally studied in the AI literature, most people may not place significant importance on whether the AI embodies deeper values, because the domain is chiefly defined by the ability to accomplish specific tasks, e.g., an AI assistant just needs to complete the requested task quickly and accurately. In these domains, people may be less likely on average to differentiate between "essential" and "superficial" characteristics, and thus may show less resistance to AI so long as its performance rivals humans (Castelo et al., 2023). However, in domains where certain values are already seen as essential (e.g., relationships, or science), there may be a stronger sense that, despite appearances, AI is "missing something essential" at its core. As a result, most people may exhibit greater resistance to adopting AI in these value-based domains.

Constraints on Generality

In Study 1, participants were recruited from Prolific, and in Study 2, from Amazon Mechanical Turk. As such, our findings are based on samples drawn from individuals registered on these online research platforms. While these platforms provide diverse and relatively representative samples of U.S. adults, the results may not fully generalize to the broader population, particularly those not active on such platforms.

References

Agarwal, S., De Freitas, J., Ragnhildstveit, A., & Morewedge, C. K. (2024). Acceptance of automated vehicles is lower for self than others. *Journal of the Association for Consumer Research*.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183-189.

Bergner, A. S., Hildebrand, C., & Häubl, G. (2023). Machine Talk: How Verbal Embodiment in Conversational AI Shapes Consumer–Brand Relationships. *Journal of Consumer Research*, ucad014.

Bertacchini, F., Bilotta, E., & Pantano, P. (2017). Shopping with a robotic companion. *Computers in Human Behavior*, *77*, 382-395.

Bower, A. H., & Steyvers, M. (2021). Perceptions of AI engaging in human expression. *Scientific Reports*, *11*(1), 1-7.

Brakus, J. J., Schmitt, B. H., & Zarantonello, L. (2009). Brand experience: what is it? How is it measured? Does it affect loyalty? *Journal of Marketing*, *73*(3), 52-68.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Castelo, N. (2023). Understanding and improving consumer reactions to service bots. *Journal of Consumer Research*.

Castelo, N., Boegershausen, J., Hildebrand, C., & Henkel, A. P. (2023). Understanding and improving consumer reactions to service bots. *Journal of Consumer Research*.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809-825.

Clark, M. S., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of personality and social psychology*, *37*(1), 12.

Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, *86*(1), 132-148.

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, *7*, 1845–1854.

De Freitas, J., Castelo, N., Uğuralp, A. K., & Oğuz-Uğuralp, Z. (2024). Lessons from an app update at Replika AI: Identity stability in human-AI relationships. *Harvard Business School Working Paper*, *25-018*.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences*, *21*(9), 634-636.

De Freitas, J., & Cohen, I. G. (2025). Disclosure, humanizing, and contextual vulnerability of generative AI chatbots. *New England Journal of Medicine AI*.

De Freitas, J., Oğuz-Uğuralp, Z., Uğuralp, A. K., & Puntoni, S. (2024). *AI Companions Reduce Loneliness*. https://ssrn.com/abstract=4893097

De Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, *10*, 3061.

Earp, B., Do, D., & Knobe, J. (2021). The Ordinary Concept of True Love.

Earp, B. D., Mann, S. P., Aboy, M., Awad, E., Betzler, M., Botes, M., Calcott, R., Caraccio, M., Chater, N., Coeckelbergh, M., Constantinescu, M., Dabbagh, H., Devlin, K., Ding, X., Dranseika, V., Everett, J. A. C., Fan, R., Feroz, F., Francis, K. B., . . . Clark, M. S. (2025). Relational Norms for Human-AI Cooperation. https://doi.org/10.48550/arxiv.2502.12102

Esch, F. R., Langner, T., Schmitt, B. H., & Geus, P. (2006). Are brands forever? How brand knowledge and relationships affect current and future purchases. *Journal of Product & Brand Management*, *15*(2), 98-105.

Fournier, S. (1998). Consumers and their brands: Developing relationship theory in consumer research. *Journal of Consumer Research*, *24*(4), 343-373.

Gao, G. (2001). Intimacy, passion, and commitment in Chinese and US American romantic relationships. *International Journal of Intercultural Relations*, *25*(3), 329-342.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619-619.

Guo, C., Dweck, C. S., & Markman, E. M. (2021). Gender categories as dual-character concepts? *Cognitive Science*, *45*(5), e12954.

Hayes, A. F. (2012). *PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]*. Retrieved from http://www.afhayes.com/public/process2012.pdf

Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of Marketing*, *70*(4), 19-36.

Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P., & Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, *32*(3), 401-414.

Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Research on aging*, *26*(6), 655-672.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242-257.

Laursen, B., & Hartup, W. W. (2002). The origins of reciprocity and social exchange in friendships. *New Directions for Child and Adolescent Development*, *2002*(95), 27-40.

Liu, Y., Mittal, A., Yang, D., & Bruckman, A. (2022, 2022). Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. New York, NY, USA.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91-108.

Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing*, *85*(2), 14-32.

Muniz Jr, A. M., & O'guinn, T. C. (2001). Brand community. *Journal of Consumer Research*, *27*(4), 412-432.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81-103.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, *146*(2), 165.

Reich, T., Kaju, A., & Maglio, S. J. (2022). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285-302.
Rusbult, C. E. (2003). Interdependence in close relationships. *Blackwell handbook of social psychology: Interpersonal processes*, 357-387.
Scott, C. A. (1976). The effects of trial and incentives on repeat purchase behavior. *Journal of Marketing Research*, *13*(3), 263-269.
Waytz, A., & Norton, M. I. (2014). Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. *Emotion*, *14*(2), 434-444.
Yin, Y., Jia, N., & Wakslak, C. J. (2024). AI can help people feel heard, but an AI label diminishes this impact. *Proceedings of the National Academy of Sciences*, *121*(14), e2319112121.