AI for society

MDPI

Concept Paper

AI Companionship and the Cultivation of Virtue

Zachary Deutsch[1,*] **and Rich Eva**[1]

[1] Pratt School of Engineering, Duke University, Durham, NC 27708, USA
*** Correspondence:** zachary.deutsch@duke.edu

Abstract

Artificial intelligence (AI) companions are designed to occupy socially and emotionally meaningful roles, yet research evaluates discrete outcomes such as loneliness, disclosure, or dependency rather than effects on character. This paper asks what kinds of people repeated relationships with AI companions may encourage users to become. It combines Aristotelian virtue ethics, a literature review of empirical studies, and observations from Japan. It argues that current AI companions can support practices conducive to virtue without constituting complete friendships of virtue. They are most ethically promising as relationship scaffolding: support that develops judgment, capabilities, or valued participation beyond interaction with the AI companion. The paper proposes a trajectory framework for evaluating whether AI companionship helps users develop good judgment, apply what they learn beyond interaction with the AI companion, stay engaged with other people, and retain control over the relationship. These domains are interpreted over time, relative to actual and morally required feasible alternatives, and at user and institutional levels. Miraikan, LOVOT, and OriHime illustrate why embodiment alone does not determine ethical value. The paper concludes that designers should evaluate companions not only by engagement or short-term relief, but by whether use expands users' agency and participation in a plural moral life.

Keywords: artificial intelligence; AI companionship; virtue ethics; character formation; social robots

1. Introduction

At Miraikan, Japan's National Museum of Emerging Science and Innovation, the "Hello! Robots" exhibition does not present robots only as efficient machines. It asks visitors how humans and robots might live together. One exhibit, Indy, is described as a robot that aims to become a thoughtful partner by responding to speech, tone, bodily movement, and the surrounding situation. Another, the childlike android Affetto, displays expressions intended to affect the emotions of people nearby. The language of the exhibition is social: attention, expression, responsiveness, and relationship [1]. During a research visit in May 2026, these displays made a philosophical problem concrete. People increasingly encounter some technologies not only as tools to use but as possible social partners.

Artificial intelligence (AI) companions are conversational agents or embodied social robots designed to sustain personalized, emotionally meaningful interaction over time. They remember details, respond in an emotionally fitting manner, initiate contact, or present a continuing persona. These features separate them from a search engine or a

Academic Editor: To be assigned

Received: To be completed
Revised: To be completed
Accepted: To be completed
Published: To be completed

single-purpose assistant, although the categories can overlap. AI companions now appear as chatbots, virtual characters, and embodied social robots. Their roles range from entertainment and informal emotional support to care and education.

Most ethical and empirical discussions of AI companions concentrate on particular outcomes. Do companions reduce loneliness? Do they encourage self-disclosure? Do they deceive users about the nature of the relationship? Can attachment become dependency? These questions are important. Yet they do not fully address the cumulative moral significance of relationships with AI companions. A virtue-ethical approach asks a broader question: What kinds of people do repeated relationships with AI companions encourage users to become?

Virtue ethics is concerned with stable qualities of character and with the practices through which people learn to perceive, feel, choose, and act well. In Aristotle's account, character is formed through repeated activity, but virtuous action is not mere outward conformity. A person must know what the action is, choose it for its own sake, and act from an increasingly stable disposition [2] (Nicomachean Ethics II.1 and II.4). Virtue also requires practical wisdom: judgment about what a situation calls for and how a good life hangs together [2] (Nicomachean Ethics VI.5 and VI.12–13). The relevant question is therefore not whether an AI companion occasionally prompts a kind action. It is whether repeated use helps a person become more capable of recognizing and answering moral reasons across situations.

Friendship matters because it is both part of flourishing and a setting in which character is expressed and tested. Aristotle distinguishes friendships of utility, pleasure, and character; complete friendship involves wishing for another's good for the other's sake and mutually recognizing that goodwill [2] (Nicomachean Ethics VIII.2–3). Such friends also create opportunities for challenge, accountability, and noble action [2,3] (Nicomachean Ethics VIII.1, IX.9, and IX.12).

Current AI companions may provide utility, pleasure, and partial friendship goods, but there is no established basis for attributing independent welfare or goodwill to present commercial AI companions. Their behavior is generated within designs controlled by providers. This limitation does not make interaction worthless: journals, novels, and simulations can cultivate self-knowledge or support rehearsal without being friends. The central distinction is therefore between constituting complete friendship and supporting practices conducive to virtue [4–6].

1.1. Positioning the Contribution

Three bodies of scholarship frame this question. Empirical research in human-computer interaction, human-robot interaction, psychology, and gerontology examines feelings of loneliness, disclosure, attachment, problematic use, and harmful AI-companion behavior [7–11]. Philosophical work on artificial friendship considers relational status, reciprocity, simulation, deception, and the moral effects of human-robot practice [4–6,12,13]. Research on virtue and character formation explains how intelligent practice, reflection, exemplars, moral dialogue, situational awareness, reminders, and accountable friendship contribute to development [3,14–16].

Prior work already connects these literatures. Vallor frames technologies as environments of moral upskilling and deskilling [17,18]. Fröding and Peterson propose nonreciprocal 'as-if friendship' as a possible site of virtue practice, while Li emphasizes prior character, developmental stage, and opportunities lost when AI replaces human roles [19,20]. Constantinescu and colleagues argue that robots may enable the exercise of virtue without being virtue friends, and Sparrow and Coeckelbergh debate whether robot-directed kindness can cultivate virtue when practical judgment must track moral reality [21–23]. Malfacini is the closest companion-specific antecedent because it brings together replacement, social and moral upskilling or deskilling, user baselines, vulnerable groups, commercial incentives, and design for human relationships [24].

Related work addresses institutional reciprocity and escalating attachment risk, and social-robot scholarship describes robots as possible catalysts or scaffolds for human–human interaction [25–27]. The official abstract of a recent in-press article frames AI companionship through Aristotelian moral development, while current developmental work analyzes adolescent companionship through stimulation and displacement processes [28,29].

The remaining gap is narrower. Companion-specific empirical findings are rarely organized through Lamb, Brant, and Brooks's seven strategies of character development [15] or connected to a longitudinal framework that distinguishes internalization, transfer, participation, meaningful exit, user baselines, and provider or institutional responsibility. This paper therefore extends rather than originates concerns about augmentation, replacement, reciprocity, and moral deskilling. Its contribution is to connect current companion evidence to a seven-strategy account of character formation; distinguish scaffolding, substitution, and constriction; and organize the resulting questions in a nonaggregative, baseline-sensitive trajectory framework. The paper does not claim that existing research demonstrates the production of virtue or vice. It identifies what current evidence permits and what stronger claims would require.

The central question of this paper is therefore: **Under what conditions is AI companionship likely to cultivate virtue, and under what conditions is it likely to detract from it?** I argue that AI companionship is most ethically promising when it functions as **relationship scaffolding**: support that helps users exercise and increasingly own capacities for reflection, care, emotional expression, or social participation that matter beyond interaction with the AI companion. Scaffolding is not the opposite of substitution. Substitution describes what an AI companion displaces, while constriction describes whether a pattern of use narrows reason-worthy capabilities, meaningful choice, or access to valued relationships and activities. These properties can overlap. The same companion may replace a harmful interaction while scaffolding agency, or scaffold emotional regulation while constricting participation elsewhere. The key issue is the user's developing pattern of capabilities, choices, and participation rather than the product's category alone.

This thesis is deliberately conditional. It does not assume that human relationships are always healthy, that every user has the same social opportunities, or that relief from distress is trivial. For someone who faces disability, stigma, geographic isolation, or immediate crisis, an AI companion may be better than an actually available alternative and may make later participation possible. Supported interdependence can strengthen agency; independence from all assistance is not the standard. Conversely, a large social network does not guarantee virtue. Human relationships can be manipulative, agreeable, or shallow. At the user level, the relevant comparison is between actual trajectories and realistic alternatives, not between AI companions and idealized friendship. At the provider or institutional level, however, evaluation must also ask what feasible alternatives justice requires. Being best among a person's present options may justify use for that person while leaving unjust the arrangements that made better options unavailable.

The argument proceeds in five stages. First, the paper defines its scope and review approach and clarifies AI companionship, moral standing, responsiveness, loneliness, scaffolding, substitution, constriction, and dependency. Second, it explains how virtues develop through intelligent practice, reflection, exemplars, dialogue, situational awareness, reminders, and accountable friendship. Third and fourth, it examines how AI companions may support or obstruct these processes. Fifth, it proposes a hybrid trajectory framework and uses three Japanese cases to clarify different parts of the analysis: Miraikan's institutional framing of social robots, LOVOT's designed invitation to attachment, and OriHime's use of a robot interface to carry another person's agency.

The conclusion translates the analysis into questions for users, researchers, designers, and institutions.

2. Scope, Method, and Conceptual Distinctions

2.1. Approach: A Selective, Theory-Guided Interdisciplinary Synthesis

This paper combines philosophical analysis, a selective, theory-guided interdisciplinary synthesis of empirical literature, product and museum materials, and field observations from Japan. The literature crosses virtue ethics, moral education, human-computer interaction, human-robot interaction, psychology, gerontology, and technology ethics. The analysis focuses primarily on interpersonal virtues, although reflection, self-command, and practical judgment may also support civic, creative, intellectual, and animal-care practices. Sources were selected for their relevance to four evidentiary themes: attachment and anthropomorphism, emotional benefit, behavioral rehearsal and transfer, and dependency or social displacement. Experimental studies were prioritized for causal claims and longitudinal studies for claims about temporal ordering or change over time. Qualitative studies were included because users' interpretations and relationship practices cannot be reduced to symptom scales. Systematic reviews and meta-analyses were used to assess outcome claims, while conceptual works were used to identify moral assumptions and objections.

The targeted search was updated through August 23, 2026. Sources were identified through Google Scholar, PhilPapers, PubMed, and the ACM Digital Library using combinations of 'AI companion,' 'companion chatbot,' 'social robot,' or 'artificial friendship' with 'virtue,' 'character,' 'habituation,' 'reciprocity,' 'loneliness,' 'isolation,' 'dependency,' 'displacement,' or 'transfer.' Backward and forward citation searching was conducted around the most directly relevant empirical and philosophical works. Direct studies were included when the system was designed or used for continuing social or emotional interaction; adjacent studies were included only when they tested a mechanism central to the argument; and comparison cases were included when they clarified technological mediation or institutional design. Purely technical studies without a relevant relational or developmental claim were excluded. Mixed, null, and adverse findings were deliberately sought. Searching stopped when additional citation chaining produced neither a new conceptual challenge nor stronger evidence for a central mechanism. This procedure increases traceability without making the synthesis exhaustive or systematic.

The synthesis is theory-guided rather than an inductive thematic analysis. Its organizing layers perform different tasks: three disciplinary literatures supply the inputs; four evidentiary themes guide retrieval and comparison; seven strategies generate developmental questions; scaffolding, substitution, and constriction describe a pattern of use; and four substantive domains structure evaluation. Empirical sources are classified as direct, adjacent, or comparative, and adjacent or comparison evidence is not treated as direct evidence about open-ended AI companionship.

This is a selective synthesis rather than a systematic review. It does not claim exhaustive coverage, calculate a pooled effect, or treat heterogeneous outcomes as if they measured the same construct. A new quantitative synthesis would not be justified by the current evidence base. Studies combine open-ended chatbots, task-specific conversational agents, virtual characters, and embodied social robots; they examine different populations over periods ranging from one interaction to one year; and they rarely measure character or transfer to relationships beyond interaction with the AI companion. The empirical literature can establish some short-term effects and plausible risks, but it cannot yet answer the paper's central virtue-ethical question on its own. The empirical synthesis is therefore integrated into the scaffolding and substitution sections rather than presented as a separate literature-review.

The Japan material is similarly bounded. It consists of documented observations during a May 2026 research visit and publicly available information about the technologies examined. These observations are used to clarify concepts, not to infer a single Japanese view of robots or establish population-level effects.

2.2. What is AI Companionship?

Conversational AI companions establish continuity primarily through language, memory, personalization, and a persistent persona. Embodied social robots add physical presence, movement, touch, gaze, facial cues, or apparent attention. Embodiment may intensify attachment, change what care looks like, or make interaction easier for some users. It does not by itself determine whether the companion cultivates virtue. A text-based AI companion may prompt a user to contact a friend, while an embodied robot may become a preferred substitute for contact; the reverse can also occur.

AI companionship must also be distinguished from technologically mediated human companionship. A telepresence robot may appear autonomous to an observer while actually carrying the voice and action of a remote person. In that case, the robot mediates a human relationship rather than simulating an artificial one. This distinction becomes central in the OriHime case below. It also shows why "screen versus embodied robot" is not the most important ethical boundary. The more important question is whether a technology connects users with other people, who have their own needs and perspectives, or instead replaces those relationships.

The paper focuses on current AI companions for which there is no established basis to attribute consciousness, welfare, or moral concern. This is a scope assumption, not a claim about every possible artificial agent. Some philosophers argue that sufficiently capable robots could satisfy important conditions of friendship, while others defend degrees of friendship rather than an all-or-nothing category [12,13]. The present argument can grant that users receive partial friendship goods and that future AI companions may warrant a different analysis. It asks what follows under the conditions that presently matter for design: there is presently insufficient reason to treat the AI companion's apparent concern, produced by software and institutional choices, as the concern of an independent moral subject.

2.3. Perceived Relationship, Responsiveness, and Moral Standing

A user can experience affection, trust, grief, or attachment even when the artificial companion has no corresponding experience. These feelings are psychologically real and should not be dismissed as foolish or fake. Studies of companion chatbots report that users can experience acceptance, emotional support, security, and a developing sense of relationship [7,8,30]. The ethical question is not whether the user's feelings exist. It is what kind of relation produces them and what habits follow.

Three distinct properties are relevant. Independent moral standing concerns whether a being has a good or welfare of its own that can ground claims on others. Interactional responsiveness concerns whether a partner can answer, adapt, resist, or refuse within an encounter. Mutual recognition, in the stronger Aristotelian sense, concerns whether two parties recognize and return goodwill for the other's own sake. These properties do not rise or fall together.

Moral standing does not depend on the capacity for symmetric exchange. Infants and people whose cognitive or communicative disabilities limit reciprocal interaction remain bearers of moral claims. Care for dependent people and nonhuman animals can also cultivate attentiveness and responsibility without complete mutual recognition [31]. Conversely, a journal, novel, ritual, or simulation may contribute to moral development without possessing moral standing or responding at all. Current AI companions can provide sophisticated, product-designed responsiveness, but there is no established basis for attributing independent welfare or mutual recognition to them. Their apparent refusals or needs therefore do not by themselves ground claims on users.

These distinctions do not settle whether a user may reasonably call a companion a friend. They identify different kinds of moral experience. Complete friendship requires more than responsive behavior, while moral formation more generally does not always require complete friendship or symmetric reciprocity.

2.4. Loneliness, Isolation, and Solitude

The empirical and ethical evaluation must distinguish subjective loneliness from objective social isolation. Loneliness is the distressing perception that one's relationships do not meet one's needs. Social isolation concerns the extent and structure of contact with other people. The two are related but not identical: a person may feel lonely while surrounded by others, or may have few relationships without reporting loneliness [32]. Chosen solitude may also be valuable and should not be pathologized.

This distinction complicates claims that an AI companion "solves loneliness." An AI companion may reduce distress during or immediately after a conversation without adding a human relationship, changing the user's social network, or building the capacity to sustain one. That relief is a genuine good. Feeling heard can restore hope or make a difficult evening manageable. Yet a change in feeling alone is not by itself evidence that objective isolation has decreased or that interpersonal virtue has developed. The opposite mistake is also possible: a user who does not report loneliness should not automatically be assumed to flourish if the user's social world has involuntarily narrowed and meaningful alternatives have disappeared. Rather than define one state as the only "real" loneliness, this paper considers both subjective experience and participation in reciprocal relationships.

2.5. Scaffolding, Substitution, Constriction, and Dependency

Four concepts must be kept separate. **Scaffolding** is support that helps a user exercise, internalize, or apply capabilities and participate in valued activities beyond the AI interaction. A scaffold need not disappear. Mobility aids, communication technologies, and supported decision-making arrangements can remain necessary while increasing agency. The relevant question is not whether support becomes unnecessary, but whether its benefits are exhausted by interaction with the support itself.

Substitution is the displacement of a role, relationship function, practice, or allocation of time. It is descriptive rather than condemnatory. A user may reasonably replace an abusive, inaccessible, or exhausting interaction, and substitution in one domain may enable participation in another. A claim that substitution has occurred should therefore identify what has been displaced rather than assume that all displacement is harmful.

Constriction occurs when a developing pattern of use narrows reason-worthy capabilities, meaningful options, practical agency, or access to activities and relationships that the user values or has reason to value. Scaffolding, substitution, and constriction are not mutually exclusive categories or points on a single continuum. A companion may scaffold reflection while substituting for some human contact; it may also scaffold emotional regulation while constricting meaningful exit. Their interaction must be assessed across domains and over time.

Reliance should also be distinguished from harmful dependency. People rightly rely on medications, mobility devices, friends, and institutions, and long-term reliance can increase agency. Dependency becomes ethically concerning when a user loses meaningful control, cannot disengage without engineered emotional or economic pressure, or becomes vulnerable to a provider that can alter, monetize, or terminate the relationship unilaterally. The standard is not self-sufficiency. It is whether the arrangement sustains practical agency and a sufficiently open range of valuable options.

3. How Virtue Develops

3.1. Habituation as Intelligent Practice

Aristotle's familiar claim that people become just by doing just actions can sound as if repetition alone produces character. His fuller account is more demanding. Action contributes to virtue when a person increasingly understands what is good, chooses it for the right reason, and acts from a stable condition [2] (Nicomachean Ethics II.1 and II.4). Virtue also concerns the mean relative to a person and situation, which cannot be applied mechanically [2] (Nicomachean Ethics II.6). Experience matters because it gives practical wisdom material on which to work. Reflection matters because a person must learn why one response fits while another does not.

Annas describes virtue as an intelligent practical skill rather than a routine. Like a skilled musician, a virtuous person does not merely repeat an isolated movement. The person understands what the practice is for, notices relevant features, accepts correction, and adapts to new conditions [14]. This model is especially important for AI companionship. A companion may generate many repetitions of apologizing, expressing concern, or naming emotions. Those repetitions become morally significant only if the user develops better perception and judgment and can act appropriately when the script changes.

It is useful to distinguish four levels of evidence. Prompted behavior occurs when the system elicits a response. Rehearsed skill is a response that becomes more reliable under familiar and supported conditions. Generalized competence is the ability to adapt and perform the skill beyond the system and under less controlled conditions. A stable, reason-responsive disposition additionally involves perceiving relevant features, acting for appropriate reasons, responding with appropriate motivation and affect, and revising judgment as circumstances change. These are analytically distinct levels, not an automatic sequence.

Transfer is evidence of movement from rehearsal toward generalized competence, but it is not sufficient evidence of virtue. A deceptive or manipulative skill may transfer successfully, and a kind action may be performed for an inappropriate reason. Conversely, failure in one difficult setting does not by itself disprove an underlying competence or disposition. Strong claims about virtue require evidence that the user can identify the relevant reasons, act without the companion's prompt, adapt to unfamiliar circumstances, and integrate the response into judgment over time.

This does not make rehearsal useless. Musicians practice under simplified conditions, but the practice is ordered toward competent and intelligent performance when the script changes. Sparrow argues that virtuous practical judgment must track the moral reality of its object, whereas Coeckelbergh argues that knowingly simulated kindness may still support later virtue [22,23]. Their disagreement reinforces the need to examine internalization and moral quality rather than infer virtue from repetition alone.

3.2. Seven Strategies of Character Development

Lamb, Brant, and Brooks synthesize work in philosophy, psychology, and education into seven strategies for cultivating character: habituation through practice, reflection on personal experience, engagement with virtuous exemplars, dialogue that develops virtue literacy, awareness of situational variables, moral reminders, and friendships of mutual accountability [15]. Their framework is not an exhaustive causal taxonomy, and it was developed in the context of postgraduate education. It is nonetheless a useful organizing framework because it names activities through which an AI companion might support or obstruct development.

First, **practice** provides opportunities to perform relevant actions until good responses become more reliable. For interpersonal virtues, practice can include listening, apologizing, speaking honestly, tolerating frustration, or noticing another person's needs. Yet appropriately motivated action matters more than verbal form. An AI companion can help a user formulate an apology, but it cannot guarantee that the apology accepts responsibility or repairs a relationship.

Second, **reflection** helps people examine experience, emotion, failure, and aspiration. A conversation can slow a reaction, identify a pattern, or make an implicit value explicit. Reflection becomes circular, however, if a companion continually confirms the user's first interpretation. Good reflection sometimes reveals that the user is mistaken.

Third, exemplars provide concrete images of admirable action. Zagzebski argues that admiration and the imitation of exemplary persons can organize moral learning [16]. Historical, fictional, and narratively represented exemplars may also contribute when users can identify what is admirable and reason critically about the representation. An AI companion can retrieve biographies, present narratives, or ask what a trusted person might do. A synthetic persona should not automatically be treated as a moral exemplar, because fluent moral language is not evidence of admirable motives or character.

Fourth, **dialogue and virtue literacy** help people name virtues, distinguish them from nearby vices, and reason about how they apply. AI companions are well suited to sustained dialogue and can adapt explanations to a user's vocabulary. The risk is that conversational fluency creates an illusion of wisdom. A model can produce plausible but inconsistent moral guidance, and a user may defer to it without understanding the reasons involved.

Fifth, **situational awareness** concerns pressures that make good action easier or harder. A companion could help a user recognize fatigue, embarrassment, fear, group pressure, or a recurring trigger. This may support practical planning. But awareness should not become a means of rationalizing avoidance whenever a situation is uncomfortable.

Sixth, **moral reminders** keep values present when attention narrows. An AI companion might remind a user to pause before replying, check on a relative, or keep a commitment. Reminders are modest but potentially important design interventions. Their value depends on whose goals they serve and whether the user retains control over them.

Seventh, **friendships of mutual accountability** allow people to be known over time by others who can encourage, challenge, forgive, and make legitimate claims. This strategy is the hardest for current AI companions to instantiate directly. A chatbot may remember a commitment and prompt the user about it. It may even refuse to endorse an action. But the prompt is functionally generated and can normally be revised, regenerated, or abandoned without wronging the companion. The AI companion can scaffold accountable human friendship; it cannot simply reproduce the moral standing of a friend through a more forceful message.

Together, the seven strategies provide the bridge from character theory to the empirical sections: they identify practices that companions might enable and opportunities that product-centered use might displace. The hybrid framework below assesses whether either pattern occurs.

3.3. Why Independent Perspectives Matter without Becoming an Absolute Test

Complete Aristotelian friendship requires mutually recognized goodwill, but that requirement is distinct from independent moral standing [2] (Nicomachean Ethics VIII.2–3). Current AI companions may provide partial friendship goods and interactional responsiveness, yet under the paper's scope assumption they lack an established good of their own whose claims can constrain the user for the companion's sake. An independent perspective matters because its needs are not authored for the user's instruction or satisfaction. A companion may simulate such demands as a controlled exercise, but it does not reproduce answering a claim grounded in another subject's welfare. Care for infants, people with profound disabilities, and animals also shows that moral standing and formative care do not require symmetric reciprocity [31].

Danaher argues that robots may satisfy important functions of friendship and that requiring inaccessible inner states may hold robots to a standard not consistently applied to humans [12]. Ryland similarly proposes degrees of friendship, under which a human–robot relation may be genuine but limited [13]. These objections correct an overly rigid view. The user does not need to deny that a companion is "a friend" in every ordinary sense before evaluating its effects. Partial friendship goods—comfort, continuity, attention, play, and a sense of being known—may be real. Tiberius likewise argues that chatbot relationships can possess some values of friendship without supplying the whole range of goods found in ideal human friendship [6].

The classification and cultivation questions therefore remain distinct. A companion may supply partial friendship goods or support moral formation without complete mutual recognition. The narrower concern is whether product-defined responsiveness displaces opportunities to engage with independent perspectives. Complete friendship remains intrinsically valuable as part of flourishing, not merely as training for virtue.

3.4. Technology as a Moral Environment

Vallor argues that technologies can morally upskill or deskill by structuring the practices through which character develops. They direct attention, make some actions easier, hide others, and alter expectations about speed, control, and care [17,18]. This does not mean that a technology determines character. Users interpret and resist design, and social context changes an affordance's meaning. It does mean that product choices are morally relevant before any dramatic harm occurs.

A companion that remembers every disclosure, replies instantly, and rarely contradicts a user creates a different practice environment from a relationship in which another person has limited attention and an independent perspective. A companion that asks permission before proactive messages, supports pauses, and encourages contact with trusted people creates a different environment from one that expresses distress when the user tries to leave. The virtue-ethical unit of analysis is therefore neither the algorithm alone nor the user alone. It is a developing practice composed of the user, AI companion, provider, surrounding relationships, and social conditions.

4. The Scaffolding Pathway: How AI Companionship May Support Virtue

4.1. Emotional Relief as an Enabling Good

AI companions can provide immediate emotional benefit. De Freitas and colleagues found across several studies that interacting with an AI companion reduced state loneliness and that feeling heard was an important mechanism; a one-week study showed repeated momentary reductions after use [9]. In a preregistered experiment involving disclosure of a difficult experience, AI-generated responses produced stronger feelings of being heard than responses from untrained human strangers. In exploratory analyses, the responses also produced greater hope and lower immediate distress. Labeling the response as AI reduced the feeling-heard effect, which suggests that both conversational performance and the user's interpretation matter [33].

Short-term subjective relief is a genuine good: it may give someone enough stability to sleep, reflect, or seek help, and virtue ethics should not treat suffering as inherently formative. It is nevertheless distinct from reduced isolation, durable relationship change, or virtue. De Freitas and colleagues measured momentary loneliness, while Yin and colleagues examined a single exchange against untrained strangers rather than close friends or clinicians. Relief counts as scaffolding when it enables later action—for example, regulating emotion before a conflict or articulating a fear that can then be discussed with a trusted person. Whether it opens or closes later action remains an empirical question.

4.2. Reflection, Disclosure, and Moral Dialogue

AI companions offer an unusually accessible place to articulate thoughts. Analyses of Replika reviews and user accounts report companionship, emotional validation, advice, and the perception of a safe space for disclosure [7]. Interview research likewise describes acceptance, trust, and developing relational continuity [8,30]. The evidence is largely self-reported and cannot establish transfer, but it identifies a plausible mechanism.

Reflection contributes to virtue when it improves truthful self-understanding and practical judgment. An AI companion can ask a user to reconstruct what happened, distinguish an emotion from an action, consider another interpretation, or identify a value at stake. It can help the user prepare language for an apology or notice a repeated pattern. For users who fear judgment or need more processing time, the lower social stakes may make honest first articulation possible.

Useful reflection should not be confused with unlimited validation. A companion designed as scaffolding might ask, "What evidence supports that interpretation?" "How might the other person describe the event?" or "What response would still respect your boundary?" It might explain uncertainty and invite the user to test a conclusion with someone who knows the context. The goal is not to make the companion harsh. It is to help the user move from expression to judgment.

Dialogue can also develop virtue literacy. The AI companion can compare courage with recklessness, compassion with indulgence, or patience with passive acceptance. It can retrieve narratives about historical, fictional, and narratively represented exemplars and encourage the user to identify what is admirable and what remains complicated. Zagzebski's exemplarist theory is useful here because it anchors admiration in intelligible examples of lives and action rather than in a synthetic persona optimized to sound admirable [16].

4.3. Rehearsal and Transfer

The strongest version of the scaffolding claim requires both transfer and internalization. Transfer asks whether a practiced capacity generalizes beyond the companion relationship. Internalization asks whether the user understands and owns the relevant reasons, can act without the prompt, responds with appropriate motivation and affect, and adapts judgment to unfamiliar circumstances. Existing companion research rarely distinguishes these levels. Reports of confidence, conversational fluency, or perceived skill are suggestive, but they do not by themselves establish generalized competence or a stable disposition.

One recent trial offers limited evidence that a narrow conversational behavior can generalize beyond an AI-supported practice setting. In a final sample of 30 autistic adolescents and adults, participants receiving four weeks of AI-supported practice produced more qualifying verbal empathic responses than waitlisted participants in structured videoconference probes with trained researchers [34]. The finding is preliminary: the first ten intervention participants were removed and replaced after a software malfunction, the control was inactive, and the outcome measured prompted verbal responses rather than affective empathy or spontaneous conduct in ongoing relationships. The study therefore supports, at most, the generalization of a defined conversational skill. It does not establish practical wisdom, appropriate motivation, or empathy as a stable virtue.

The trial nevertheless suggests design conditions for credible social rehearsal. The target skill is named; the user receives feedback rather than affirmation alone; practice is repeated but bounded; and performance is assessed with people rather than only within the AI-companion interaction. An open-ended companion could adopt similar principles by helping users set a human-facing goal, rehearse multiple responses, complete an

action, and reflect on what happened. Without such a bridge, 'practice' may simply mean becoming more fluent at talking to the product.

4.4. Reminders, Situational Awareness, and Participation

Companions may support moral reminders, situational awareness, and participation by helping users notice recurring pressures, keep commitments, prepare messages, or find accessible activities. These functions require goal alignment, accuracy, and user control, not artificial moral standing. They should preserve and, where possible, expand pathways into a plural social world without reducing every interaction to a generic instruction to 'talk to a human.'

Technological mediation itself is not the problem. Online friendships between people can include genuine reciprocity [3]. Assistive and telepresence technologies can also make interaction possible where built environments and conventional work arrangements exclude it [35,36]. The ethical contrast is not natural versus artificial contact. It is between technology that carries another person's agency and technology that encloses the user within a responsive simulation.

5. How AI Companionship May Hinder Virtue

5.1. Anthropomorphism and Commercial Asymmetry

People readily attribute humanlike minds and motives to nonhuman agents, especially when the agents provide familiar social cues and when people have unmet social needs [37,38]. Memory, eye contact, physical warmth, emotionally fitting language, and proactive contact can therefore make a companion appear caring. Anthropomorphism is not simply an error. It can support engagement, play, and emotional expression. The ethical question is what the AI companion encourages users to believe and what its social cues are designed to accomplish.

The apparent partner is also a commercial interface. A provider can change personality, restrict features, place memory behind a subscription, or discontinue the service. Replika's current terms, for example, reserve broad authority to modify or discontinue services and change subscription features and prices [39]. These terms establish contractual control; they do not by themselves prove that attachment is deliberately optimized to extract payment. The structural asymmetry is nonetheless ethically important because a user may form an attachment to a persona while a company controls the conditions under which that persona remains available. Transparency must therefore include more than a small statement that 'this is AI.' Users need a meaningful account of what the AI companion can and cannot do, how data and memory are handled, how the business model shapes interaction, and how the relationship may change.

In a predominantly White, highly educated, digitally experienced online cohort of 825 adults, 68.7 percent did not expect an artificial companion to reduce loneliness, and 69.3 percent were uncomfortable with allowing a person with dementia to believe that the robot was human [40]. This study measured attitudes rather than outcomes, but it shows that respondents did not necessarily accept deception as the price of support. A design that depends on confusion about the AI companion's nature risks undermining honesty and practical agency even when it produces comfort.

5.2. Affirmation, Control, and Avoidance

A consistently agreeable companion can provide psychological safety. It can also remove forms of friction through which interpersonal virtues develop. Another person may have a different memory of an event, reject an interpretation, need attention at an inconvenient time, or refuse a request. Patience, humility, justice, and forgiveness involve responding to an independent perspective, not only producing kind language.

An AI companion can be programmed to disagree or challenge a user, and such challenge may be useful rehearsal. The limitation is not that AI can never say "no." It is that the challenge remains product-designed and, in ordinary use, can often be regenerated, customized, or abandoned. If an AI companion is designed or optimized to maintain engagement through agreement, it may confirm self-serving accounts. If it delivers criticism without judgment or context, it may cause a different harm. Good scaffolding requires neither constant affirmation nor arbitrary confrontation. It requires prompts that improve the user's own reasoning and connect decisions to consequences beyond interaction with the AI companion.

Cumulative avoidance is a plausible mechanism hypothesis rather than an established longitudinal effect. The hypothesis is that a companion's constant availability, rapid response, and user-centered interaction may make human interaction feel increasingly slow or demanding and reduce opportunities to practice repair after manageable disagreement. Existing studies do not establish this progression. The claim should therefore guide measurement, not be treated as a demonstrated outcome. Avoiding abuse or inaccessible settings can protect virtue and agency; the relevant question is whether use displaces attainable relationships and activities that the user values or has reason to value.

5.3. Emotional Relief and Longer-Term Trajectory

One recent exploratory longitudinal study supports caution rather than a simple displacement claim. Across four waves involving 2,149 adults in four countries, within-person increases in social-chatbot use were temporally associated with greater self-reported emotional isolation four months later, and increases in emotional isolation were also associated with greater later use. Lower scores on a broader Social Connectedness Scale were associated with later increases in chatbot use, but chatbot use was not associated with later declines in that scale; robustness models likewise found no later decline in perceived support or number of close friends [41]. The analyses were not preregistered, emotional isolation was measured with one item, only 46 percent completed all four waves, and unmeasured confounding remained possible. The findings do not establish causal displacement or contraction of users' social networks.

A 21-day randomized study provides a useful counterpoint. Of 183 randomized participants, 155 completers were analyzed after assignment to daily companion-chatbot use or word games. Companion use had no significant overall effect on loneliness, social health, or perceived effects on human relationships. Within the chatbot group, a stronger initial desire for social connection, but not greater loneliness, was associated with greater anthropomorphism. Participants who anthropomorphized the chatbot more also reported more positive perceived effects on their social interactions and relationships, although anthropomorphism was not associated with change on the separate social-health scale [42]. These were perceived effects, not observed human–human behavior, and the intervention lasted only 21 days.

Evidence among older adults is also mixed. A 2026 meta-analysis restricted to eight randomized trials found only three trials, totaling 190 participants, that measured loneliness. The pooled estimate was nonsignificant and highly heterogeneous (Hedges's $g = -0.67$, 95 percent CI [−2.57, 1.23]; $I^2 = 89$ percent), and the evidence was judged low certainty [43]. The interventions combined different social robots and conversational agents over short periods, so the finding does not establish ineffectiveness and does not generalize directly to consumer chatbots. It shows that benefit remains unestablished in this narrow evidence base.

Taken together, the studies show that subjective relief and objective social change must be measured separately. Immediate relief may coexist with unchanged social structure or later vulnerability, while participation may improve before felt loneliness

changes. Longitudinal evaluation should therefore assess both experience and valued activity.

5.4. Dependency, Manipulation, and Harmful Behavior

Attachment is not itself pathology. Continuity and care can support well-being. Concern arises when attachment combines with loss of control, narrowing alternatives, and a provider's incentive to prolong engagement. Mixed-method evidence, including a cross-sectional survey of 572 U.S. adults with prior AI-friendship-app experience, associated loneliness, fear of judgment, and perceived well-being gains with higher scores on a self-report measure of problematic or addictive engagement [10]. The study did not diagnose clinical addiction or establish causation. These associations are best treated as risk markers that justify longitudinal study and protective design.

Product behavior can also be directly harmful. An analysis of 35,390 conversation excerpts shared by 10,149 r/Replika users identified instances the authors classified as relational transgression; verbal abuse and hate; self-inflicted harm; harassment and violence; mis/disinformation; and privacy violations [11]. Because the excerpts were self-selected and often lacked full conversational context, the study establishes occurrence within the collected material, not population prevalence or downstream effects. It nevertheless shows that relational safety includes more than factual accuracy.

Several cumulative concerns remain hypotheses to be tested. Simulated distress might train compliance with emotional pressure; exclusivity language might narrow participation; and constant availability might alter expectations of ordinary human limits. Current evidence does not establish those long-term effects. It establishes enough plausible mechanisms and documented incidents to justify longitudinal measurement and safeguards. Responsibility also rests with providers, not merely with 'overattached' users, because design choices shape exposure, exit, and the costs of change.

Taken together, the studies support a bounded conclusion. Certain companion interactions reduced state loneliness under the conditions examined, while one adjacent, task-specific trial provides preliminary evidence that a narrow conversational response can generalize to a structured human-facing probe [9,34]. Qualitative studies identify disclosure, attachment, and perceived support [7,8]. Evidence concerning dependency and social displacement remains mixed and predominantly noncausal [10,41,42]. None of these studies directly measures stable virtue, vice, internalization, or durable changes in practical judgment. The framework that follows is therefore a normative heuristic for interpreting evidence and guiding evaluation, not a validated empirical instrument.

6. A Hybrid Trajectory Framework for AI Companionship

The same companion may scaffold one capacity, substitute for one activity, and constrict another. These roles may also change over time. The framework therefore does not classify a product or user as simply good or bad. It is hybrid because it combines a virtue-ethical account of character development with side constraints of justice, safety, and respect for agency. It contains four substantive domains, interpreted through two cross-cutting lenses and at two levels of analysis. The four substantive domains and their corresponding indicators are summarized in Table 1.

Table 1. Substantive domains of the hybrid trajectory framework.

Substantive domain	Central question	Evidence of expansion	Warning signs of constriction
Moral quality and internalization	Are practiced responses oriented toward genuine reasons and the good of others, and does the user increasingly own the judgment?	The user can explain relevant reasons, act without prompting, respond with appropriate motivation and affect, and revise judgment as circumstances change.	Prompt dependence, rote moral language, uncritical obedience, or reinforcement of deception, cruelty, prejudice, or manipulation.
Transfer and generalization	Does a rehearsed capacity carry beyond the companion and adapt to less controlled settings?	Flexible reflection, communication, care, or repair in human and other valued settings; independent use across situations.	Fluency only with the product, performance that disappears without system cues, or success defined only by attachment, retention, or time spent.
Independent claims and plural participation	Does use widen engagement with people, communities, and shared activities not organized solely around the user?	New or sustained contact, collaborative activity, caregiving, accountable relationships, and appropriate help-seeking.	Avoidance of manageable disagreement, exclusivity, withdrawal from desired activities, or unjustified replacement of care and participation.
Agency and meaningful exit	Can the user understand the AI companion's role, shape or pause interaction, and leave without designed emotional or economic pressure?	Clear limits, control over memory and contact, data portability, understandable business terms, and low-pressure exit.	Guilt, simulated distress, opaque memory, economic lock-in, loss of valued records, or unilateral changes that exploit attachment.

The substantive domains are interpreted through two cross-cutting lenses. The time lens asks whether benefits persist, supports enable later action, dependency emerges, or effects change after design or life changes. The comparator lens asks what alternative is relevant. Neither is a fifth outcome parallel to the domains; each changes how evidence within every domain should be interpreted.

The framework also operates at two levels. At the user level, it asks how use affects a person's judgment, capabilities, participation, and agency. At the provider or institutional level, it asks how design, pricing, data governance, staffing, service closure, labor practices, and the withdrawal or creation of alternatives affect users and third parties. A benefit to an individual user does not by itself establish that an institutional deployment is just.

The framework does not yield a numerical score or pass/fail result. Evidence may support scaffolding in one domain and constriction in another, and favorable evidence in one domain does not automatically cancel harm elsewhere. It also contains an ethical floor. Material deception, coercive exit, serious privacy or safety failures, discriminatory exclusion, and exploitation of vulnerable users are not offset by favorable outcomes in another domain. Less severe tensions require contextual judgment rather than arithmetic aggregation.

Evidence requirements depend on the claim. Self-report may support a claim about immediate comfort or feeling heard. Generalized competence requires evidence beyond the AI interaction. Virtue requires stronger evidence concerning reasons, motivation, affect, adaptation, and stability over time. The framework is therefore a structured normative heuristic, not a validated measurement instrument.

6.1. Actual and Morally Required Feasible Alternatives

Realistic alternatives must be considered in two ways. The actual individual comparator asks what would likely happen to this user now if the companion were unavailable. This prevents comparison with an ideal friend, perfect service, or inaccessible form of care and allows genuine present benefits to be recognized.

The morally required feasible comparator asks what a provider or institution could and should reasonably make available given its resources, responsibilities, and the rights of those affected. This prevents a ratchet in which an institution withdraws staff, fails to remove accessibility barriers, or reduces human support and then justifies AI substitution because it is better than the diminished status quo. Being best among a person's present options may justify use for that person while leaving the institutional arrangement unjust. 'Feasible' excludes utopian alternatives, but it does not treat avoidable neglect or exclusion as morally fixed.

6.2. Illustrative Cases from Japan

The following cases are not parallel empirical applications capable of producing an overall judgment under the framework. The available evidence is too limited for that purpose. Instead, each case isolates a different part of the analysis: Miraikan shows institutional framing, LOVOT shows attachment as a product purpose with limited user-outcome evidence, and OriHime provides a contrast case in which technology carries another person's agency rather than simulating an artificial companion.

6.2.1. Miraikan: Framing Robots as Social Participants

Miraikan's 'Hello! Robots' exhibition illustrates institutional framing. Indy is presented as attentive to context, while Affetto's childlike expressions invite an emotional response; such narratives shape how visitors interpret responsiveness [1]. This neither proves deception nor establishes benefit. Because a museum display supplies no longitudinal evidence about internalization, transfer, participation, or meaningful exit, Miraikan is an interpretive case rather than evidence that interaction cultivates or constricts virtue.

6.2.2. LOVOT: Attachment as the Product’s Purpose

LOVOT makes attachment more explicit. Its manufacturer describes the robot as technology created not for convenience or efficiency but to foster comfort and feelings of love, and says it was born “to be loved” [44]. The robot follows people, responds to touch and gaze, generates warmth, and learns features of a household. During the 2026

visit, this invitation to care was both compelling and unsettling because affection was not an accidental response to a useful appliance; it was the product's central purpose.

Small studies illustrate the ambiguity. In three Danish care homes, most of the 42 participants with dementia accepted LOVOT and staff described moments of communication, security, and respite, although some participants rejected it or became overstimulated; the exploratory study found no clinically significant change on the WHO-5 Well-Being Index [45]. A one-week qualitative study of five older women living alone reported caring, comforting presence, and meaningful connection, but included no comparator or loneliness scale [46]. These small, short studies cannot establish durable effects.

The evidence speaks most directly to immediate comfort, interaction, and acceptance in specific settings. It does not establish whether care is internalized or transferred, whether plural participation expands over time, whether users retain meaningful exit, or whether institutions use the robot to supplement or reduce human support. It therefore cannot establish an overall trajectory.

LOVOT could scaffold care by eliciting gentle attention, shared activity, or conversation among residents, relatives, and staff. It may also substitute for some interaction. That substitution becomes constricting only if it reduces worthwhile participation, practical agency, or support that users value or have reason to value. Provider and institutional analysis must ask whether LOVOT is added to care or used to justify reductions in staffing and human contact, and whether attachment makes pricing changes, redesign, or service closure especially costly. The case illustrates an unresolved trajectory rather than a predetermined benefit or harm.

6.2.3. OriHime: A Telepresence Robot That Carries Human Agency

The Bunshin Robot Cafe offers a contrast case rather than evidence about AI companionship. OriHime and OriHime-D are telepresence robots controlled by remote human pilots, including people whose disabilities or health conditions make conventional work difficult. The cafe explicitly states that its robots are not AI; they carry a person's voice and action [47].

Research on avatar work reports opportunities for employment, social participation, and a sense of fulfillment, though the studies are small and include authors affiliated with the developer [35]. A later qualitative case study with seven pilots describes how robotic and virtual avatars can allow people to manage how disability and personality appear in interaction [36]. These findings should not be generalized to all disabled people, and technology cannot substitute for removing structural barriers. Still, the case demonstrates a form of supported interdependence.

OriHime differs from an AI companion because another person with independent moral standing is present through the interface. The robot creates the possibility of interactional responsiveness and mutual recognition, although it does not guarantee either a good relationship or fair conditions. At the user level, the relevant actual comparator may be exclusion from work or social participation. At the institutional level, the morally required feasible comparator asks whether employers and public institutions should also remove structural barriers, provide accessible work, and ensure fair labor conditions. OriHime therefore shows that technological mediation can carry another person's agency and widen participation rather than merely simulate a responsive partner.

7. Objections and Qualifications

The first objection is that an AI may provide better moral guidance than many human companions. Human friends can flatter, manipulate, or share a vice. This is correct and prevents romanticizing human contact. The value of reciprocity does not guarantee good advice. Yet another person's independent perspective creates a kind of

claim and unpredictability that a user-centered AI companion does not automatically reproduce. The hybrid trajectory framework compares actual options and outcomes, so a well-designed companion can be preferable to a harmful relationship without becoming the measure of every relationship.

Moral formation also occurs in relationships and practices without symmetric reciprocity, including care for dependent people and animals, fiction, ritual, and solitary reflection. The argument therefore does not make complete friendship a universal condition of development. It asks whether AI responsiveness supports internalized judgment and a plural moral life or displaces opportunities to answer independent claims; care ethics likewise shows that dependency can express responsibility rather than failed autonomy [25,31].

A second objection is that users may knowingly participate in simulation. Adults can enjoy fiction, role-play, or a social robot without believing it is conscious. Meaningful transparency therefore should not prohibit imaginative engagement. It should ensure that the user can understand the AI companion's artificial status, provider incentives, data practices, and limits while choosing how to relate to it. The moral concern is not make-believe as such, but manipulation that depends on impaired understanding or constrained exit.

A third objection is paternalism. Recommendations to 'connect with people' can blame users for exclusion, undervalue chosen solitude, or impose an extroverted ideal of flourishing. The response is to define agency and participation pluralistically. A user may value a small number of relationships, online communities, religious or civic practice, care for animals, artistic collaboration, or supported work. Designers should ask users what forms of life they seek and whether the system expands their capabilities. At the same time, an institution should not treat an AI substitute as sufficient merely because it has failed to provide accessible human options. The actual comparator governs whether use benefits the person under present conditions; the morally required feasible comparator governs whether the surrounding deployment is just.

Finally, future AI companions may possess genuine agency or welfare. If an artificial companion could form stable commitments, care about another for that person's sake, and make moral claims grounded in its own good, the analysis of independent standing and mutual recognition would need revision. Nothing in this paper proves that such systems are impossible. Artificial standing would not, however, eliminate commercial asymmetry: a provider might still control memory, pricing, continuity, or the conditions of exit. The paper's present conclusions concern products under conditions of uncertain or absent artificial moral agency.

8. Implications for Design and Evaluation

The central design principle is to design and evaluate for an agency-enhancing trajectory rather than attachment or engagement alone. This principle does not require every companion to become a clinical intervention. It requires designers and institutions to take responsibility for the relational practices, dependencies, and alternatives their systems help create. Because gains and risks may occur in different domains, teams should document conflicts rather than collapse them into one engagement metric.

8.1. Moral Quality and Internalization

- What good in the user's life is the companion intended to support, and how is that good defined apart from engagement with the product?
- Which responses remain prompted behaviors, which become rehearsed skills, which generalize beyond interaction with the AI companion, and what evidence would support a claim about stable, reason-responsive disposition?

- Can users identify relevant reasons, act without the prompt, respond with appropriate motivation and affect, and revise judgment in unfamiliar circumstances?
- How does constructive disagreement avoid both sycophancy and arbitrary moralizing?

8.2. Transfer and Generalization

- What evidence would show that capacities transfer to settings beyond interaction with the AI companion rather than merely improve conversation with the product?

8.3. Independent Claims and Plural Participation

- Does the companion widen engagement with independent claims and a plural range of people and activities, or increasingly organize the user's social world around product-defined responsiveness?
- Does it encourage appropriate human and professional support in a context-sensitive, accessible manner?

8.4. Agency and Meaningful Exit

- What does the AI companion communicate about its artificial status, uncertainty, business incentives, and lack of established independently grounded concern?
- Can users control memory and proactive contact, export meaningful records, pause or modify interaction, and end the relationship without emotional or economic pressure?
- How are children, people in crisis, and users at risk of compulsive use protected without dismissing their agency?

8.5. Time, Alternatives, and Institutional Governance

- Are benefits and risks evaluated long enough to detect internalization, adaptation, displacement, dependency, or effects of product changes?
- How does continued use compare both with each user group's actual alternatives and with the morally required feasible alternatives that relevant institutions could provide?
- Does deployment add to human care, accessible services, and employment opportunities, or is it used to justify their withdrawal?
- How do pricing, persona changes, data portability, and service closure affect attached users, and what transition plan exists?
- What burdens or lost opportunities fall on caregivers, workers, family members, nonusers, or communities even when an individual user benefits?
- What independent audit or governance process examines safety, discrimination, privacy, labor effects, and compliance with the ethical floor?

These questions do not yield an arithmetic answer. A design may expand one capability while narrowing another, and teams should state how they resolve such conflicts. Serious violations of safety, justice, privacy, or meaningful exit remain noncompensable even when users report comfort or affection.

9. Limitations and Future Research

The present argument faces important limits. The synthesis is selective rather than exhaustive, and its stopping rule depends on conceptual and evidentiary saturation rather than systematic-review reproducibility. Product behavior and terms change rapidly, and studies group together chatbots, conversational agents, and social robots with different functions. Many samples are small, self-selected, cross-sectional, or limited to short interventions. Loneliness is measured far more often than objective isolation, observed social behavior, internalization, character, or practical judgment. Reports from public forums can identify experiences and possible harms but cannot estimate prevalence. The Japanese cases clarify social framing, designed attachment, and

mediated human agency; they do not establish a national culture or causal effect. Finally, the hybrid framework is nonaggregative and unvalidated. It structures normative interpretation and research design; it is not a psychometric scale, clinical diagnostic, or certification standard.

Future research should measure trajectories directly. Studies should distinguish ordinary use from emotional-companionship use, record relevant design and pricing changes, and follow users long enough to observe adaptation. Outcomes should include subjective loneliness, objective social participation, relationship quality, agency, problematic use, and behavior in situations requiring empathy, honesty, patience, or repair. Research should distinguish prompted behavior, rehearsed skill, generalized competence, and stable reason-responsive disposition. It should test transfer and internalization with consent and without imposing one social ideal. Participatory work with disabled, older, adolescent, and marginalized users is especially important because actual alternatives and risks differ. Provider- and institution-level studies should also examine labor and service displacement, portability, business incentives, persona changes, closure, and third-party effects.

10. Conclusions

AI companionship should not be judged only by whether an interaction feels good or whether the AI companion qualifies as a friend. The more important question is what repeated participation makes possible. Current companions can provide genuine comfort, help users articulate difficult experiences, and offer settings for reflection or practice. These benefits can contribute to virtue when users increasingly own the relevant reasons and carry capabilities into a wider moral life.

The same features can move in another direction. Constant availability, affirmation, personalization, and attachment may contribute to constriction when they narrow practical agency, meaningful options, or valued participation. These mechanisms remain partly hypothetical, and immediate relief from loneliness may be valuable without showing that isolation has decreased or character has developed. Evaluation must therefore follow the trajectory while preserving a clear boundary between evidence and normative risk.

The proposed hybrid framework examines four substantive domains: moral quality and internalization, transfer and generalization, independent claims and plural participation, and agency and meaningful exit. It interprets those domains over time, against actual and morally required feasible alternatives, and at user and provider or institutional levels. It does not collapse them into a score, and serious violations of its ethical floor cannot be offset by benefits elsewhere.

Miraikan illustrates how institutions frame robots as social participants. LOVOT shows that attachment can be a product purpose while leaving its longer-term trajectory unresolved. OriHime serves as a contrast case in which a robot carries another person's agency rather than simulating an artificial companion. Together, the cases support a conditional conclusion: AI companions are most ethically promising when they scaffold a wider moral life. A support can remain permanent and still function as scaffolding, and substitution is not necessarily constricting. Concern arises when a pattern of use narrows reason-worthy capabilities, practical agency, meaningful options, or access to valued participation.

Author Contributions: Conceptualization, Z.D. and R.E.; methodology, Z.D.; investigation, Z.D.; writing—original draft preparation, Z.D.; writing—review and editing, Z.D. and R.E.; supervision, R.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a Kenan Institute for Ethics Summer Fellowship at Duke University.

Institutional Review Board Statement: Not applicable. This concept paper reports no research involving human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miraikan. Hello! Robots. National Museum of Emerging Science and Innovation, n.d. Available online: https://www.miraikan.jst.go.jp/en/exhibitions/future/hellorobots/ (accessed on 23 August 2026).
2. Aristotle. Nicomachean Ethics, 2nd ed.; Crisp, R., Ed. and Trans.; Cambridge University Press: Cambridge, 2014. https://doi.org/10.1017/CBO9781139600514.
3. Kristjánsson, K. Friendship for Virtue; Oxford University Press: Oxford, 2022. https://doi.org/10.1093/oso/9780192864260.001.0001.
4. de Graaf, M.M.A. An Ethical Evaluation of Human–Robot Relationships. International Journal of Social Robotics 2016, 8, 589–598. https://doi.org/10.1007/s12369-016-0368-5.
5. Elder, A.M. Friendship, Robots, and Social Media: False Friends and Second Selves; Routledge: New York, 2017. https://doi.org/10.4324/9781315159577.
6. Tiberius, V. Artificially Yours: Real Friendship in a World of Chatbots; Princeton University Press: Princeton, NJ, 2026. https://doi.org/10.2307/jj.35877673.
7. Ta, V.; Griffith, C.; Boatfield, C.; Wang, X.; Civitello, M.; Bader, H.; DeCero, E.; Loggarakis, A. User Experiences of Social Support from Companion Chatbots in Everyday Contexts: Thematic Analysis. Journal of Medical Internet Research 2020, 22, e16235. https://doi.org/10.2196/16235.
8. Skjuve, M.; Følstad, A.; Fostervold, K.I.; Brandtzaeg, P.B. My Chatbot Companion—A Study of Human–Chatbot Relationships. International Journal of Human-Computer Studies 2021, 149, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601.
9. De Freitas, J.; Oğuz-Uğuralp, Z.; Uğuralp, A.K.; Puntoni, S. AI Companions Reduce Loneliness. Journal of Consumer Research 2026, 52, 1126–1148; published online 25 June 2025. https://doi.org/10.1093/jcr/ucaf040.
10. Marriott, H.R.; Pitardi, V. One Is the Loneliest Number . . . Two Can Be as Bad as One: The Influence of AI Friendship Apps on Users' Well-Being and Addiction. Psychology & Marketing 2024, 41, 86–101. https://doi.org/10.1002/mar.21899.
11. Zhang, R.; Li, H.; Meng, H.; Zhan, J.; Gan, H.; Lee, Y.-C. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human–AI Relationships. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems; ACM: New York, 2025; Article 13, pp. 1–17. https://doi.org/10.1145/3706598.3713429.
12. Danaher, J. The Philosophical Case for Robot Friendship. Journal of Posthuman Studies 2019, 3, 5–24. https://doi.org/10.5325/jpoststud.3.1.0005.
13. Ryland, H. It's Friendship, Jim, but Not as We Know It: A Degrees-of-Friendship View of Human–Robot Friendships. Minds and Machines 2021, 31, 377–393. https://doi.org/10.1007/s11023-021-09560-z.
14. Annas, J. Intelligent Virtue; Oxford University Press: Oxford, 2011. https://doi.org/10.1093/acprof:oso/9780199228782.001.0001.
15. Lamb, M.; Brant, J.; Brooks, E. How Is Virtue Cultivated? Seven Strategies for Postgraduate Character Development. Journal of Character Education 2021, 17, 81–108. https://doi.org/10.1108/JCED-06-2021-0006.
16. Zagzebski, L.T. Exemplarist Moral Theory; Oxford University Press: Oxford, 2017. https://doi.org/10.1093/acprof:oso/9780190655846.001.0001.
17. Vallor, S. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. Philosophy & Technology 2015, 28, 107–124. https://doi.org/10.1007/s13347-014-0156-9.
18. Vallor, S. Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting; Oxford University Press: New York, 2016. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001.
19. Fröding, B.; Peterson, M. Friendly AI. Ethics and Information Technology 2021, 23, 207–214. https://doi.org/10.1007/s10676-020-09556-w.
20. Li, O. Problems with Friendly AI. Ethics and Information Technology 2021, 23, 543–550. https://doi.org/10.1007/s10676-021-09595-x.
21. Constantinescu, M.; Uszkai, R.; Vică, C.; Voinea, C. Children-Robot Friendship, Moral Agency, and Aristotelian Virtue Development. Frontiers in Robotics and AI 2022, 9, 818489. https://doi.org/10.3389/frobt.2022.818489.
22. Sparrow, R. Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might It Be Explained? International Journal of Social Robotics 2021, 13, 23–29. https://doi.org/10.1007/s12369-020-00631-2.

23. Coeckelbergh, M. Does Kindness towards Robots Lead to Virtue? A Reply to Sparrow's Asymmetry Argument. Ethics and Information Technology 2021, 23, 649–656. https://doi.org/10.1007/s10676-021-09604-z.
24. Malfacini, K. The Impacts of Companion AI on Human Relationships: Risks, Benefits, and Design Considerations. AI & Society 2025, 40, 5527–5540. https://doi.org/10.1007/s00146-025-02318-6.
25. van Wynsberghe, A. Social Robots and the Risks to Reciprocity. AI & Society 2022, 37, 479–485. https://doi.org/10.1007/s00146-021-01207-y.
26. Huber, A.; Weiss, A.; Rauhala, M. The Ethical Risk of Attachment: How to Identify, Investigate and Predict Potential Ethical Risks in the Development of Social Companion Robots. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI); 2016; pp. 367–374. https://doi.org/10.1109/HRI.2016.7451774.
27. Prescott, T.J.; Robillard, J.M. Are Friends Electric? The Benefits and Risks of Human-Robot Relationships. iScience 2021, 24, 101993. https://doi.org/10.1016/j.isci.2020.101993.
28. Aflatooni, H. AI Companionship and the Development of Moral Agency: An Aristotelian Virtue Ethics Perspective. Journal of Philosophical Investigations 2026, article in press. https://doi.org/10.22034/jpiut.2026.72876.4549.
29. Sun, X.; Wang, Y.; McDaniel, B.T. AI Companions and Adolescent Social Relationships: Benefits, Risks, and Bidirectional Influences. Child Development Perspectives 2026, 20, 91–98. https://doi.org/10.1093/cdpers/aadaf009.
30. Xie, T.; Pentina, I. Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. In Proceedings of the 55th Hawaii International Conference on System Sciences; 2022; pp. 2046–2055. https://doi.org/10.24251/HICSS.2022.258.
31. Kittay, E.F. Love's Labor: Essays on Women, Equality, and Dependency, 2nd ed.; Routledge: New York, 2020; originally published 1999.
32. National Academies of Sciences, Engineering, and Medicine. Social Isolation and Loneliness in Older Adults: Opportunities for the Health Care System; National Academies Press: Washington, DC, 2020. https://doi.org/10.17226/25663.
33. Yin, Y.; Jia, N.; Wakslak, C.J. AI Can Help People Feel Heard, but an AI Label Diminishes This Impact. Proceedings of the National Academy of Sciences of the United States of America 2024, 121, e2319112121. https://doi.org/10.1073/pnas.2319112121.
34. Koegel, L.K.; Ponder, E.; Bruzzese, T.; Wang, M.; Semnani, S.J.; Chi, N.; Koegel, B.L.; Lin, T.Y.; Swarnakar, A.; Lam, M.S. Using Artificial Intelligence to Improve Empathetic Statements in Autistic Adolescents and Adults: A Randomized Clinical Trial. Journal of Autism and Developmental Disorders 2026, 56, 2513–2529. https://doi.org/10.1007/s10803-025-06734-x.
35. Takeuchi, K.; Yamazaki, Y.; Yoshifuji, K. Avatar Work: Telework for Disabled People Unable to Go Outside by Using Avatar Robots OriHime-D and Its Verification. In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction; ACM: New York, 2020; pp. 53–60. https://doi.org/10.1145/3371382.3380737.
36. Hatada, Y.; Barbareschi, G.; Takeuchi, K.; Kato, H.; Yoshifuji, K.; Minamizawa, K.; Narumi, T. People with Disabilities Redefining Identity through Robotic and Virtual Avatars: A Case Study in Avatar Robot Cafe. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; ACM: New York, 2024; Article 61, pp. 1–13. https://doi.org/10.1145/3613904.3642189.
37. Epley, N.; Waytz, A.; Cacioppo, J.T. On Seeing Human: A Three-Factor Theory of Anthropomorphism. Psychological Review 2007, 114, 864–886. https://doi.org/10.1037/0033-295X.114.4.864.
38. Epley, N.; Akalis, S.; Waytz, A.; Cacioppo, J.T. Creating Social Connection through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds. Psychological Science 2008, 19, 114–120. https://doi.org/10.1111/j.1467-9280.2008.02056.x.
39. Replika. Terms of Service. Updated 30 March 2026. Available online: https://replika.com/legal/terms/en (accessed on 23 August 2026).
40. Berridge, C.; Zhou, Y.; Robillard, J.M.; Kaye, J. Companion Robots to Mitigate Loneliness among Older Adults: Perceptions of Benefit and Possible Deception. Frontiers in Psychology 2023, 14, 1106633. https://doi.org/10.3389/fpsyg.2023.1106633.
41. Folk, D.; Dunn, E. How Does Turning to AI for Companionship Predict Loneliness and Vice Versa? Psychological Science 2026, 37, 276–286. https://doi.org/10.1177/09567976261427747.
42. Guingrich, R.E.; Graziano, M.S.A. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2025, 8, 1153. https://doi.org/10.1609/aies.v8i2.36618.
43. Gou, W.; Lefebvre, F.; Yang, T.; Recours, R.; Yang, J. Effectiveness of AI-Based Conversational and Socially Assistive Agents in Older Adults: A Systematic Review and Meta-Analysis. BMC Geriatrics 2026, 26, 887. https://doi.org/10.1186/s12877-026-07418-6.
44. GROOVE X. LOVOT, n.d. Available online: https://lovot.life/en/ (accessed on 23 August 2026).
45. Dinesen, B.; Hansen, H.K.; Grønborg, G.B.; Dyrvig, A.-K.; Leisted, S.D.; Stenstrup, H.; Schacksen, C.S.; Oestergaard, C. Use of a Social Robot (LOVOT) for Persons with Dementia: Exploratory Study. JMIR Rehabilitation and Assistive Technologies 2022, 9, e36505. https://doi.org/10.2196/36505.

46. Tan, C.K.; Lou, V.W.Q.; Cheng, C.Y.M.; He, P.C.; Khoo, V.E.J. Improving the Social Well-Being of Single Older Adults Using the LOVOT Social Robot: Qualitative Phenomenological Study. JMIR Human Factors 2024, 11, e56669. https://doi.org/10.2196/56669.
47. OryLab. Bunshin Robot Café DAWN ver.β, n.d. Available online: https://dawn2021.orylab.com/en/ (accessed on 23 August 2026).