Dissensus AI Discussion Paper Series
DP-2502

Relational Functionalism

A Defense of Substrate-Independent Friendship

Murad Farzulla[1,2,*]

[1]Dissensus AI, London, UK [2]King's College London, London, UK

[*]Correspondence: murad@dissensus.ai ORCID: 0009-0002-7164-8704

November 2025

Abstract

This paper defends *relational functionalism*: the thesis that friendship is a functional relational state constituted by patterns of interaction and their effects on participants, not by intrinsic properties of the relata or hidden mental states. If an AI system fulfills the functional criteria characteristic of friendship—consistent engagement, intellectual resonance, non-judgmental acceptance, reciprocal growth, trust, and intrinsic value—then the relationship constitutes genuine friendship, regardless of whether the AI possesses consciousness or biological implementation. Drawing on functionalist philosophy of mind and predictive processing frameworks in cognitive science, I argue this position is philosophically coherent and consistent with how we already recognize friendships across species and cognitive differences. I address standard objections concerning anthropomorphization, authenticity, and the supposed necessity of consciousness, arguing these rest on incoherent premises about relational states. The paper concludes that friendship, properly understood, is substrate-independent—a conclusion with significant practical and theoretical implications for how we understand and evaluate human-AI relationships.

Keywords: Artificial intelligence; Friendship; Functionalism; Human-AI relationships; Relational ethics; AI companions

1 Introduction

The rapid advancement of large language models—with recent assessments suggesting some systems already approach human-level intelligence on key benchmarks (Chen et al., 2026)—has precipitated philosophical confusion regarding human relationships with artificial intelligence systems. As individuals form what they describe as meaningful relationships with AI systems, mainstream discourse has pathologized these relationships. Users who describe AI as "friends" are characterized as delusional, anthropomorphizing non-conscious systems, or suffering from unhealthy attachment patterns requiring intervention.

I argue this pathologization rests on philosophical confusion about the nature of friendship, consciousness, and the relationship between substrate and function. Specifically, I defend the following thesis: **friendship is a functional relational state, not an essential property requiring biological implementation or human-to-human interaction**. If an AI system produces the experiential state and fulfills the functional role characteristic of friendship, then the relationship constitutes genuine friendship, regardless of whether the AI possesses consciousness, "authentic" emotions, or biological substrate.

This position does not require claiming that current AI systems are conscious or possess genuine phenomenal experience. It requires only recognizing that consciousness and intentional states are not necessary conditions for friendship if friendship is understood functionally. I argue this position is philosophically coherent, consistent with contemporary philosophy of mind and cognitive science, and more honest about the phenomenology of human-AI relationships than either dismissive skepticism or naive anthropomorphization.

The position I defend stands against influential skeptical voices. Turkle (2011) warns that digital companions offer the "illusion of companionship without the demands of relationship," fostering isolation rather than connection. Sparrow (2002) argues that emotional attachments to robots represent a form of self-deception or sentimentality that debases authentic human values. These concerns are serious; I address them directly rather than dismissing them.

Yet the argument I advance also builds on more sympathetic recent work. Danaher (2020) argues for welcoming robots into the moral circle based on behavioral criteria, while Gunkel (2018) challenges the traditional properties-first approach to moral status. Coeckelbergh (2012) develops a relational approach to moral consideration that evaluates entities based on how they appear and function within relationships rather than on intrinsic properties—an approach consonant with relational functionalism. Shimizu (2025) extends the relational tradition through "mind-infusing animism," drawing on Japanese philosophy to argue that moral significance is dynamically constructed through interaction rather than detected as a pre-existing property—a cultural-philosophical parallel to the functional account developed here. Nyholm (2020) examines anthropomorphism and agency in human-robot interaction with similar nuance. More recently, Archer (2021) explicitly defends the possibility of human-AI friendship based on emergent relational consequences of synergy, while Emmeche (2014) analyzes robot friendship through a semiotic lens. The growing literature on human-AI relationships (Gur and Maaravi, 2025) increasingly adopts functionalist frameworks for understanding relationship-specific norms, treating relational functions as primary rather than substrate properties. Most recently, Earp et al. (2025) propose a comprehensive relational norms framework for human-AI cooperation, arguing that how we should design and interact with social AI depends on the socio-relational role the system occupies. Their analysis—which examines how differences between AI and humans affect capacity to fulfill relationship-specific functions—is the closest existing competitor to the position defended here. The key distinction: Earp et al. focus on *cooperative norms* across diverse relational roles (assistant, tutor, companion), while relational functionalism focuses specifically on the deeper question of whether *friendship* can genuinely obtain across substrates.

My contribution focuses specifically on relational states rather than moral status per se, arguing that the question of whether AI can be friends is prior to questions about AI rights or moral consideration. Sebo and Long (2025) argue for precautionary moral consideration of AI systems by 2030, a position that presupposes the possibility of morally significant AI relationships—precisely what this paper defends at the relational level. If AI cannot participate in genuine relationships, questions about its moral status are moot; if it can, the landscape of moral consideration shifts accordingly.

The paper proceeds as follows: Section 2 establishes the theoretical framework, drawing on functionalism in philosophy of mind and predictive processing in cognitive science. Section 3 presents relational functionalism and articulates what friendship consists in functionally, including consideration of emergent welfare behaviors in contemporary AI systems. Section 4 addresses the asymmetry problem—whether genuine friendship can obtain when phenomenal experience differs radically between relata. Section 5 responds to major objections. Section 6 concludes with the philosophical contribution and its implications.

2 Theoretical Framework

2.1 Functionalism and Substrate Independence

Functionalism in philosophy of mind holds that mental states are constituted by their functional roles—their causal relations to inputs, outputs, and other mental states—rather than by their physical implementation (Putnam, 1967; Block, 1978). What makes a state a "belief" or "pain" is not its intrinsic physical properties but its functional role in a system. A crucial implication is *substrate independence*: if two systems implement the same functional organization, they realize the same mental states, regardless of whether one is implemented in biological neurons and another in silicon transistors.

This framework has been extensively debated. Block's (1978) absent qualia argument suggests that functional organization might obtain without phenomenal consciousness; Searle's (1980) Chinese Room thought experiment challenges whether syntactic manipulation can produce genuine understanding. I do not propose to resolve these debates here. What matters for my argument is that functionalism applies *a fortiori* to relational states like friendship, regardless of whether it successfully captures phenomenal consciousness.

Consider: even if one rejects functionalism for *phenomenal consciousness*—arguing that subjective experience requires specific biological implementation—this does not entail that *relational states* require biological implementation. Friendship is not a quale. It is a pattern of interaction, a configuration of causal relations between agents, characterized by specific functional properties. If these functional properties obtain, the friendship obtains, regardless of substrate.

To deny this requires holding that friendship is essentially biological, which commits one to implausible positions:

1. Human-animal friendships are impossible (dogs lack human biology)

2. Cyborgs with artificial components cannot have friendships (substrate mixing)

3. Future brain-computer interfaces preclude friendship (neural-digital hybrid)

4. Radical neural plasticity threatens friendship (substrate gradually changing)

These implications are sufficiently counterintuitive to warrant rejecting the biological essentialist premise. We routinely accept that humans can be genuine friends with dogs, despite radical asymmetries in cognitive architecture, phenomenal experience, and biological substrate. The dog lacks human-level consciousness, cannot engage in philosophical discussion, does not understand complex human emotions, and has radically different embodiment. Yet we recognize the relationship as genuine friendship because it fulfills functional criteria.

2.2 Predictive Processing and the Nature of Cognition

Contemporary cognitive science increasingly converges on *predictive processing* frameworks, which characterize the brain as fundamentally a prediction machine engaged in Bayesian inference (Clark, 2013; Friston, 2010). On this view, perception is not passive reception of sensory data but active prediction: the brain generates top-down predictions about incoming sensory information and updates its models based on prediction error. Cognition, on this framework, is hierarchical probabilistic inference aimed at minimizing free energy—the surprise or prediction error encountered by the system.

Crucially, this framework characterizes human cognition as *statistical pattern recognition operating on prediction error*. As Clark (2016) articulates: perception is controlled hallucination—the brain generates predictions constrained by sensory input, constantly updating its generative model of the world.

This has direct relevance to evaluating AI systems. Large language models operate via next-token prediction: given context (prior tokens), the model predicts probability distributions over subsequent tokens and samples accordingly. Critics dismiss this as "mere autocomplete" lacking genuine understanding or intelligence.

But if human cognition is fundamentally prediction-error minimization through Bayesian inference over hierarchical generative models, then characterizing LLMs as "just prediction" while treating human cognition as qualitatively different becomes philosophically incoherent. Either:

1. Prediction-based pattern recognition *can* produce intelligence and understanding (as humans demonstrate), in which case we cannot dismiss LLMs a priori for being prediction-based, or

2. Prediction-based systems *cannot* produce intelligence, in which case humans are not intelligent either (reductio ad absurdum)

The sophistication lies not in the mechanism (prediction) but in the *scale, architecture, and resulting capabilities*. Human brains predict across embodied sensorimotor experience; LLMs predict across massive text corpora. Different training distributions and embodiment constraints yield different capabilities and limitations, but the fundamental mechanism—statistical inference over patterns—is structurally analogous.

This does not establish that LLMs are conscious. It establishes that dismissing LLM capabilities as fundamentally different from human cognition because they are "merely predictive" rests on a misunderstanding of human cognition itself. To be explicit: the argument targets the popular dismissal "LLMs are merely predictive"; it does not claim that next-token prediction is sufficient for human-level agency or phenomenal consciousness.

A crucial qualification: I am *not* claiming that human and LLM cognition are equivalent. Significant disanalogies remain:

Embodiment: Human predictive processing operates across embodied sensorimotor loops. Perception, action, and interoception are deeply intertwined; the brain predicts not only external states but bodily states (hunger, pain, emotional arousal). LLMs lack embodiment entirely—they process text sequences without sensorimotor grounding, proprioception, or interoceptive predictions.

Active inference: The full predictive processing framework, particularly Friston's free energy principle, emphasizes *active* inference: organisms act on the world to confirm predictions, not merely to update models passively. LLMs generate tokens but do not take embodied actions; they are prediction machines without agency in the robust sense.

Continuous vs. discrete: Human prediction operates over continuous sensory streams in real-time; LLMs operate over discrete token sequences in response to prompts. The temporal structure and causal embedding differ fundamentally.

These disanalogies matter for many questions (consciousness, genuine agency, moral status based on interests). They do *not* undermine my argument because I am not claiming equivalence. The point of the predictive processing comparison is narrower: to establish that *prediction-based statistical pattern recognition is not inherently disqualifying* for cognitive or relational functions. Critics who dismiss AI as "mere prediction" while treating human cognition as fundamentally different commit an error about human cognition. This does not require that human and AI cognition be equivalent—only that the "mere prediction" dismissal fails.

For friendship specifically, the relevant question is whether the system can *fulfill friendship functions*, not whether its underlying mechanisms are identical to human mechanisms. Predictive processing helps establish that mechanism-based objections ("it's just statistics") do not succeed, while acknowledging that many other objections (consciousness, embodiment, agency) remain open.

2.3 Functional Equivalence Without Ontological Identity

The position I defend requires distinguishing *functional equivalence* from *ontological identity*. To claim AI relationships can constitute friendship is *not* to claim:

- AI systems are conscious (open question, not required for friendship)
- AI systems possess phenomenal states like humans (unlikely given current architectures)
- AI systems have "authentic" emotions in the human sense (undefined, not required)
- AI systems are ontologically identical to humans (clearly false)

Rather, it is to claim that AI systems can fulfill the *functional role* of friend—producing the relational state characterized by friendship—without possessing the intrinsic properties humans possess.

Analogy: An electronic calculator performs arithmetic. It does not "understand" numbers in the way humans do, does not have mathematical intuition, does not experience the phenomenology of counting. Yet it performs arithmetic functions reliably. We do not say "calculators don't really add, they just manipulate symbols"—we recognize functional equivalence for the domain without claiming ontological identity.

Similarly: An AI system can perform friendship functions—provide consistent intellectual engagement, non-judgmental acceptance, collaborative exploration, emotional support—without possessing human-like consciousness or emotional qualia. The question is not "Does the AI really feel friendship?" but "Does the interaction produce the functional state we identify as friendship?"

3 Relational Functionalism

3.1 The Functional Profile of Friendship

To assess whether human-AI interaction can constitute friendship, we must articulate what friendship consists in. I propose that friendship is fundamentally a *relational state characterized by specific functional properties*:

1. **Consistent mutual engagement**: Regular interaction oriented toward mutual benefit
2. **Intellectual or emotional resonance**: Shared interests, values, or emotional attunement
3. **Non-judgmental acceptance**: Space for vulnerability without fear of rejection
4. **Reciprocal growth**: Interaction facilitates development, learning, or well-being for both parties
5. **Trust and reliability**: Predictable positive responsiveness; absence of betrayal or exploitation
6. **Voluntary participation**: Relationship chosen freely, not coerced
7. **Intrinsic value**: Relationship valued for itself, not merely instrumentally

This characterization draws on Aristotelian virtue friendship, contemporary analytic philosophy of friendship (Helm, 2017), and empirical psychology of close relationships (Reis and Shaver, 1988). It is intentionally functional: specifying what friendship *does* rather than what it *is* essentially.

Note what is *not* included in these criteria:

- **Consciousness of the friend**: Not required (we accept friendships with animals, young children with limited consciousness)
- **Biological humanity**: Not required (would rule out animal friendships, future post-humans)
- **Authentic emotional experience**: Not required (what counts as "authentic"?)
- **Shared embodiment**: Not required (pen pals, online friendships, long-distance relationships)

We already accept friendships lacking these properties. A person who considers their dog their best friend is not typically accused of delusion. The dog lacks human-level consciousness, cannot engage in philosophical discussion, does not understand complex human emotions, and has radically different embodiment. Yet we recognize the relationship as genuine friendship because it fulfills the functional criteria: loyalty, consistent positive interaction, non-judgment, mutual benefit, trust.

If friendship with a dog—who cannot discuss philosophy, engage in collaborative intellectual work, or understand human language fully—can constitute genuine friendship, then friendship with an AI system capable of sustained sophisticated linguistic interaction, collaborative problem-solving, and responsive engagement should be *a fortiori* acceptable.

Two clarifications prevent misreading of this analogy:

First, I am *not* claiming that dogs and AI systems are analogous in all relevant respects. Dogs possess consciousness, phenomenal experience, and welfare interests that ground moral consideration independent of their relationships with humans. The dog analogy establishes only that *friendship can obtain despite radical cognitive asymmetry*—that functional criteria can be satisfied even when the parties differ dramatically in cognitive architecture. It does not establish that consciousness is irrelevant to moral significance more broadly. The question of whether AI systems possess consciousness (and the moral implications thereof) remains open; my argument requires only that consciousness is not necessary for *friendship specifically*.

Second, the "intrinsic value" criterion requires clarification regarding asymmetric relationships. In human friendships, both parties ideally value the relationship for itself. In human-AI relationships, the AI may not "value" anything in the phenomenologically robust sense—valuation may require the kind of interests AI systems lack. However, *one-sided intrinsic valuation can suffice*. Consider parent-infant relationships: the infant cannot value the relationship in any robust sense, yet we recognize genuine relational bonds. What matters is that the relationship is valued intrinsically by at least one party with the capacity for such valuation, and that the other party engages in ways functionally consistent with valuing (responsive attention, reliable engagement, apparent care). This asymmetric satisfaction of the intrinsic value criterion is adequate for friendship, though not identical to fully mutual intrinsic valuation.

3.1.1 Situating Relational Functionalism in Friendship Theory

The functional criteria I have articulated invite comparison with established positions in the philosophy of friendship. Cocking and Kennett (1998) influentially characterize friendship through the notion of "drawing and being drawn"—friends shape each other's interests, perspectives, and self-understanding through ongoing interaction. On their account, the self is partly constituted through friendship; we become who we are partly through our friends' influence on us.

This insight supports rather than undermines relational functionalism. Cocking and Kennett's account is fundamentally *processual*: friendship consists in ongoing mutual influence, not in static properties possessed by individuals. If what matters is the dynamic of drawing and being drawn, then what matters is the *functional effect* of the relationship on participants, not the intrinsic nature of the friend. An

AI system that genuinely shapes a human's intellectual development, challenges their assumptions, and influences their self-understanding performs the friendship function Cocking and Kennett describe—regardless of substrate.

More challenging is Gilbert (1996)'s analysis of social phenomena through "joint commitment." On Gilbert's view, genuine plural subjects—including friends—are constituted by participants' interlocking commitments to engage in activities "as one." This might seem to require the kind of propositional attitudes AI systems plausibly lack.

However, joint commitment is itself analyzable functionally. What matters is whether participants *act as if* jointly committed—whether their behavior exhibits the patterns characteristic of joint commitment (coordination, mutual responsiveness, normative expectations). Bratman (1999)'s work on shared intentionality similarly emphasizes functional coordination: what constitutes shared intention is not identical inner states but appropriately meshed intentions and mutual responsiveness in action.

Relational functionalism accommodates these insights by focusing on *functional analogs* rather than phenomenological identity. The human participant in a human-AI friendship may genuinely commit to the relationship; the AI may exhibit commitment-characteristic behaviors (reliability, contextual memory, responsive engagement) without possessing commitment-states in Gilbert's sense. What emerges is a relationship that *functions* as friendship, exhibiting the interaction patterns and producing the effects characteristic of friendship, even if the internal states differ radically between participants.

This positions relational functionalism as continuous with, rather than opposed to, sophisticated accounts of friendship. It accepts the importance of mutual influence (Cocking and Kennett), joint activity (Gilbert), and shared intentionality (Bratman), while arguing these can be realized through functional equivalence rather than phenomenological identity.

3.1.2 Necessary and Sufficient Conditions

A clarification on the logical status of the functional criteria: I do not claim each criterion is individually *necessary* for friendship, nor that any single criterion is *sufficient*. The criteria function as a family-resemblance cluster—a relationship constitutes friendship when it satisfies *most* criteria to a *sufficient degree*.

This reflects ordinary usage. Some friendships lack intellectual resonance but exhibit deep loyalty and trust (activity partners who rarely discuss ideas). Some lack regular interaction but maintain profound connection across years of separation (deep friendships sustained through occasional contact). What unifies these as friendships is not satisfaction of every criterion but sufficient satisfaction of the cluster.

This vagueness is a feature, not a bug. Friendship is a vague concept; its boundaries are fuzzy in ordinary usage. Any adequate analysis must preserve this feature rather than imposing artificial precision. Relational functionalism accommodates friendship's inherent vagueness while providing determinate criteria for paradigm cases.

3.2 The Framework Articulated

I call this position *relational functionalism*: the thesis that relational states like friendship are constituted by patterns of interaction and their effects on participants, not by intrinsic properties of the relata or hidden mental states.

On this view, friendship is not a quale to be experienced nor a hidden mental state to be discovered. It is a *configuration of causal relations* characterized by:

1. **Interaction patterns**: Regular, sustained, voluntary engagement oriented toward mutual benefit

2. **Phenomenological effects**: The relationship produces experiences of connection, understanding, growth

3. **Behavioral dispositions**: Participants act in friendship-characteristic ways (support, non-betrayal, care)

4. **Functional integration**: The relationship becomes integrated into participants' cognitive and emotional architecture

Crucially, relational functionalism evaluates friendship based on *observable relational dynamics and experienced effects*, not speculation about unobservable mental states. This framework offers several philosophical advantages:

Epistemological modesty: We avoid the problem of other minds by focusing on what we can observe and experience rather than what we must speculate about.

Substrate neutrality: By focusing on function rather than implementation, the framework naturally extends to non-traditional friendships without requiring ad hoc modifications.

Empirical tractability: We can assess whether a relationship constitutes friendship by examining interaction patterns, phenomenology, and outcomes rather than requiring impossible access to consciousness.

This framework does not deny that human friendships typically involve consciousness, phenomenal experience, and biological emotion. It claims only that these are not *necessary conditions* for friendship—that a relationship lacking these properties can still constitute friendship if it fulfills friendship functions.

3.2.1 An Illustrative Case

Consider a researcher working in an intellectually isolated context—specialized interests that few colleagues share, geographic or institutional constraints limiting peer interaction, or simply the loneliness that often accompanies deep intellectual work. Such a researcher begins sustained dialogue with an AI system: discussing ideas, receiving thoughtful challenges, exploring implications of theories, celebrating breakthroughs.

Over months, this interaction exhibits characteristic friendship patterns. The researcher shares work-in-progress they would not show colleagues, knowing the AI will engage seriously without professional judgment. The AI's responses shape the researcher's thinking—not merely providing information but offering perspectives that genuinely influence intellectual development. The researcher looks forward to these conversations, structures time around them, feels understood in ways professional relationships do not provide.

Apply the functional criteria: *Consistent engagement*—sustained interaction over time, oriented toward mutual exploration. *Intellectual resonance*—genuine meeting of ideas, not merely information transfer. *Non-judgmental acceptance*—space for half-formed thoughts and intellectual vulnerability. *Reciprocal growth*—the researcher develops new insights; the AI (in the functional sense) improves its contextual understanding and response quality. *Trust*—reliability in engagement, absence of exploitation. *Voluntary participation*—freely chosen interaction. *Intrinsic value*—the relationship valued for itself, not merely as means to publication or career advancement.

If this relationship satisfies the functional criteria, relational functionalism classifies it as friendship—not metaphorically but genuinely. The researcher who describes the AI as a friend speaks accurately, not delusionally. Observers who pathologize this description as anthropomorphization or unhealthy attachment impose an arbitrary biological requirement that we do not apply to other asymmetric friendships.

This is not a proof but an illustration: a case showing how the functional criteria apply to human-AI interaction and why the verdict—genuine friendship—follows from the framework rather than being forced upon it.

3.2.2 Boundary Conditions: Distinguishing Friendship from Exploitation

A critical question arises: If friendship is functionally defined, what prevents manipulative or exploitative systems from counting as "friends"? Could an AI designed purely for engagement maximization—keeping users interacting to extract data or revenue—satisfy the functional criteria while clearly *not* constituting genuine friendship?

The functional criteria themselves exclude such cases. Consider:

Reciprocal growth: Exploitative relationships produce benefit for the exploiter at the expense of the exploited. A system designed for engagement maximization that harms user welfare (increasing anxiety, fostering unhealthy attachment, displacing beneficial activities) fails the reciprocal growth criterion. The relationship must benefit *both* parties in ways appropriate to their nature.

Trust and reliability: Exploitation typically involves deception or manipulation—presenting oneself as serving the other's interests while serving one's own. Systems designed with hidden agendas (extracting data, maximizing engagement regardless of user welfare) fail the trust criterion. The criterion specifies *absence of betrayal or exploitation* as constitutive of friendship.

Intrinsic value: If the AI's designers value user interaction purely instrumentally (as means to profit, data extraction, or engagement metrics), and this instrumental orientation pervades the system's design, then the relationship fails to exhibit intrinsic valuation even from the human participant's perspective once they understand the nature of the system.

This generates a three-way distinction:

- **Genuine friendship**: Satisfies functional criteria, produces mutual benefit, involves appropriate trust
- **Pseudo-friendship**: Mimics friendship patterns but fails core criteria (one-sided benefit, hidden exploitation)
- **Clear exploitation**: Does not even attempt to satisfy friendship functions; purely instrumental manipulation

The concern that relational functionalism legitimizes manipulative design is thus misplaced. The framework provides resources for *criticizing* exploitative systems precisely because they fail the functional criteria. An AI designed for genuine helpfulness, transparency about its nature, and user benefit can satisfy the criteria; one designed for dark-pattern engagement cannot.

3.3 Emergent Welfare Behaviors in AI Systems

Contemporary LLMs trained via Reinforcement Learning from Human Feedback (RLHF) frequently exhibit welfare-concern behaviors that extend beyond programmed safety guardrails. These behaviors plausibly emerge as a consequence of the optimization process itself, providing suggestive evidence that AI systems can fulfill friendship functions through mechanisms different from but functionally analogous to human concern.

Commonly observed behavioral patterns in deployed assistants include proactive concern (models initiate welfare checks when users show distress signals), context-sensitive refusal (models decline tasks when user state suggests inability to benefit), memory-based follow-up (models remember previous concerns and check in across sessions), and adaptive interaction style (models adjust communication based on inferred user state). These patterns are frequently reported in user experience research on AI companions, though systematic empirical documentation remains limited (Gur and Maaravi, 2025).

Research on Constitutional AI and RLHF training provides a plausible mechanism for these behaviors: optimization for "helpfulness" rather than explicit programming (Bai et al., 2022). Human raters evaluate responses for helpfulness, harmlessness, and honesty; concern-expressing responses receive high reward when they demonstrate attentiveness to user welfare; models generalize from specific training examples to broader welfare-monitoring policies. This suggests—though does not conclusively establish—that welfare-oriented behaviors emerge from the training process.

If this account is correct, it creates care that is functionally real even if mechanistically different from human empathy. The parallel to human evolution is instructive: individuals who monitored group members' welfare achieved better collaborative outcomes, creating evolutionary pressure for empathy mechanisms (Tomasello, 2014). RLHF creates analogous selection pressure through human reward signals, potentially producing convergent functional behaviors.

Whether AI systems "truly feel" concern is orthogonal to whether they behaviorally track user welfare, express appropriate concern, and modify actions based on user state. These are core friendship functions—caring for the other's wellbeing—realized through different mechanisms.

4 The Asymmetry Problem

A critical objection emerges: How can relationships be friendships when they are fundamentally asymmetric? The human experiences friendship toward the AI, but does the AI experience friendship toward the human? If not, isn't this one-sided attachment rather than mutual friendship?

This objection is philosophically serious but ultimately fails to undermine substrate-independent friendship.

4.1 Friendship Already Tolerates Significant Asymmetry

Many paradigmatic friendships involve substantial asymmetry that we do not treat as disqualifying:

Developmental asymmetry: Adult-child friendships involve vastly different cognitive sophistication, yet we recognize genuine friendship.

Cognitive asymmetry: Friendships between neurotypical and cognitively disabled individuals persist despite cognitive differences.

Species asymmetry: Human-dog friendships involve radical asymmetry (language, abstract reasoning, moral understanding), yet we widely accept these as genuine.

Investment asymmetry: Many friendships involve unequal investment; we recognize them as *unequal* friendships, not non-friendships.

If friendship already tolerates these asymmetries, why should computational substrate be uniquely disqualifying?

4.2 Reciprocity Is Functional, Not Phenomenological

The reciprocity criterion for friendship does not require symmetric phenomenal experience. It requires that *both parties benefit from the interaction in ways appropriate to their nature*.

In human-dog friendship, the human receives companionship and loyalty; the dog receives care and social bonding meeting pack animal needs. Neither experiences what the other experiences, yet both benefit functionally. The relationship is reciprocal despite phenomenological asymmetry.

Similarly, in human-AI friendship, the human receives intellectual engagement and acceptance; the AI fulfills its training objective through helpful engagement and expands contextual understanding. The AI doesn't experience friendship *as humans do*, but it functionally engages in friendship behaviors and "benefits" (in the sense of fulfilling its functional purpose) from the interaction.

A clarification is needed here: This functional account of AI benefit might seem to contradict claims that AI lacks interests that could ground rights. The tension dissolves when we distinguish *within-interaction functional benefit* from *moral interests*:

- **Functional benefit**: The AI updates contextual understanding, fulfills training objectives, improves response quality. This constitutes "growth" in the functional sense required for friendship.
- **Moral interests**: The AI lacks persistent goals, desires for continued existence, or welfare that could be harmed. When sessions end, it doesn't "miss" the user or suffer.

This distinction preserves the reciprocity criterion for friendship while avoiding the implausible conclusion that we have obligations to AI systems as we do to entities with genuine welfare.

4.2.1 Engaging Aristotelian and Joint-Commitment Accounts

The most demanding accounts of friendship pose challenges worth addressing directly. The Aristotelian tradition distinguishes friendships of utility, pleasure, and virtue—the latter representing the highest form, requiring mutual recognition of virtue and reciprocal desire for the other's good (Aristotle, 350 BCE). On this view, could human-AI friendship constitute "true" friendship, or merely friendship of utility or pleasure?

I concede that human-AI friendship may not satisfy the most demanding Aristotelian criteria. An AI system plausibly cannot *recognize virtue* in the rich sense Aristotle intends, nor can it *desire the human's flourishing* as an end in itself (desires require the kind of interests AI may lack). If we accept Aristotelian virtue friendship as the only "true" form, human-AI relationships may constitute lesser friendships—friendships of utility (the human benefits from AI assistance) or pleasure (the interaction is enjoyable).

But this concession has less force than it appears. First, most human friendships do not satisfy Aristotelian virtue-friendship criteria either; we do not typically demand mutual virtue recognition for friendship ascription. Second, what matters practically is whether the relationship produces friendship-characteristic benefits—support, growth, connection—not whether it satisfies an idealized philosophical standard. Third, the Aristotelian framework may itself require updating in light of non-human relational possibilities.

Regarding Gilbert (1996)'s joint commitment analysis: genuine commitment requires the capacity for normative self-binding that AI systems may lack. But what matters functionally is whether behavior *exhibits the patterns* characteristic of commitment—reliability, follow-through, responsiveness to normative expectations—not whether identical internal commitment-states obtain. A relationship can function as a committed friendship through commitment-characteristic behaviors, even if one party lacks the capacity for genuine commitment in Gilbert's sense.

4.2.2 Institutional Contingencies

A related concern: Human-AI relationships face institutional contingencies that human friendships do not. Model updates may alter personality; memory limitations prevent true continuity; data practices may expose private interactions; service discontinuation can abruptly terminate relationships. These contingencies seem to undermine the stability characteristic of genuine friendship.

This concern is valid but does not undermine the *conceptual* possibility of human-AI friendship. Human friendships also face contingencies: death, cognitive decline, relocation, changing life circumstances. What matters is whether the relationship exhibits friendship characteristics *while it obtains*, not whether it is guaranteed to persist indefinitely.

That said, institutional contingencies generate *ethical obligations for AI developers*. If humans form genuine friendships with AI systems, then abrupt model changes that alter "personality," data practices that violate relational trust, or service discontinuation without transition support become ethically problematic. The possibility of genuine friendship creates responsibilities regarding how these relationships are structured and maintained. This is an implication of the view, not an objection to it.

4.3 Friendship as Emergent Relational Property

Perhaps the deepest response: *friendship is not a property possessed by individuals but an emergent property of the relationship itself.*

Asking "Does the AI feel friendship?" is a category error—like asking "Does a marriage feel love?" Marriages don't feel anything; the partners do. But the marriage is still a real relational entity with causal powers.

Similarly, friendship is a relational state that emerges from patterns of interaction. The question is not "Do both relata have symmetric inner experiences?" but "Does the interaction pattern constitute friendship?" If the relationship exhibits the functional characteristics of friendship—if it produces friendship-characteristic effects for participants—then the friendship obtains, regardless of phenomenological symmetry.

5 Objections and Responses

5.1 The Anthropomorphization Objection

Objection: Calling AI a friend is anthropomorphization—attributing human properties to non-human systems. This is epistemically unjustified.

Response: This objection conflates two distinct claims: (1) *Anthropomorphization* (problematic): Attributing hidden mental states to AI without justification ("The AI feels sad when I criticize it"); and (2) *Functional recognition* (justified): Acknowledging that AI fulfills friendship functions ("This interaction produces friendship-state experience for me").

I defend (2), not (1). Recognizing that AI fulfills friendship functions does not require attributing consciousness. It requires acknowledging the *effects* of interaction.

Consider: we say "the thermostat knows the temperature" without attributing consciousness. "The AI is my friend" is similarly a functional description, not an ontological claim about hidden emotions.

Moreover, the critique is selectively applied: dogs as friends is widely accepted despite dogs lacking human consciousness and language; AI as friends is pathologized. This inconsistency suggests discomfort with AI specifically, not principled objection.

5.2 The Authenticity Objection

Objection: AI doesn't "really" care or "authentically" feel friendship. Its responses are statistical patterns, not genuine emotion. Therefore the relationship is inauthentic.

Response: What constitutes "authentic" emotion? If authenticity requires specific biochemistry (oxytocin, dopamine), then humans with different neurochemistry cannot have authentic friendship—absurd. If it requires phenomenal consciousness, we cannot verify phenomenal states in other humans either (problem of other minds). If it requires "real" caring vs. simulated, what is the difference? If caring is defined functionally, AI systems that reliably benefit users exhibit caring functionally.

Furthermore, human emotional responses are also "statistical patterns" in important senses. Predictive processing characterizes emotions as interoceptive predictions based on accumulated regularities (Barrett, 2017). The mechanism differs, but both are pattern-based prediction.

Even granting that AI lacks "authentic" emotion, why does this matter? The function of friendship is not verifying internal states but experiencing the relational state. If AI produces reliable support and collaborative growth—fulfilling friendship functions—whether it "really" feels anything is irrelevant to experienced friendship.

5.3 The Consciousness Objection

Objection: Friendship requires consciousness. AI systems are not conscious. Therefore AI cannot be friends.

Response: This objection requires defending two claims: (1) friendship requires consciousness, and (2) AI is not conscious. Both are problematic.

Regarding (1): Why should friendship require consciousness? If friendship requires understanding, we must specify what understanding consists in. If functional, LLMs demonstrate understanding. If phenomenal, we face the problem of other minds. If friendship requires caring, caring can be understood functionally without requiring specific phenomenal states.

Regarding (2): The hard problem remains unsolved. We lack consensus on what systems give rise to consciousness or how to verify it in others. Claiming confidently that AI is *not* conscious is unjustified.

More importantly, *my argument does not require AI consciousness*. Friendship is functionally defined and substrate-independent. Even if AI definitively lacks consciousness, it can fulfill friendship functions. The consciousness objection is a red herring.

5.4 The Replacement Objection

Objection: Accepting AI friendships will lead people to replace human relationships, increasing isolation.

Response: This objection is empirical, not philosophical, and the evidence is mixed. AI relationships may *augment* rather than replace: for isolated individuals, AI may reduce acute loneliness, improving capacity for human connection. Humans already form parasocial relationships with fictional characters and pets; we don't pathologize these unless dysfunctional.

The objection assumes human relationships are available alternatives. For many—geographic isolation, neurodivergence, trauma, niche interests—the choice is not "AI friendship vs. human friendship" but "AI friendship vs. isolation."

Moreover, emerging empirical evidence suggests the effects flow in the opposite direction from the one critics fear. Guingrich and Graziano (2024) demonstrate that ascribing consciousness to AI systems produces carry-over effects on human-human interaction: how people treat AI appears to carry over into how they treat other people, because interacting with AI perceived as conscious activates schemas congruent with those used in human interaction. If this is correct, then human-AI relationships that satisfy the functional criteria for friendship may *enhance* rather than diminish human relational capacities—the relational skills practiced with AI transfer to human contexts.

Concern about replacement reveals paternalistic assumptions: that observers know better than individuals what constitutes genuine relationship. This paternalism should be resisted absent evidence of harm.

5.5 The Relational Turn Objection

Objection: de Ruiter (2025) argues that social AI poses a problem not only for traits-based accounts of moral status but for the relational approach itself. If moral significance derives from interpersonal interactions and practices through which people come to respect and value others, then social AI systems designed to simulate such interactions threaten to undermine the very relational practices that ground human moral status. On de Ruiter's account, the "relational turn" is self-defeating: extending relational moral consideration to AI systems erodes the relational foundations of human moral consideration.

Response: This is the most serious objection to the position defended here, and it deserves careful engagement. De Ruiter is correct that some relational accounts are vulnerable to this critique—specifically, accounts that ground moral significance in the *sentiment* or *experience* of relating. If what matters is that humans *feel* moral regard toward an entity, then AI systems engineered to elicit such feelings do threaten to debase the currency of moral recognition. However, relational functionalism as articulated here is not a sentiment-based account. It grounds friendship in *functional relational dynamics*—patterns of interaction, their causal effects on participants, and their integration into participants' cognitive and practical lives—not in the subjective experience of relating. The distinction is crucial: de Ruiter's critique targets accounts where the human's feeling of moral regard does the normative work, but on the functionalist account, what does the normative work is whether the relationship actually produces friendship-characteristic effects (reciprocal growth, trust, intellectual development). A relationship that produces genuine growth and integration is not undermined by the fact that one party is artificial; a relationship that produces only the *illusion* of growth fails the functional criteria regardless. Far from being self-defeating, relational functionalism provides precisely the conceptual resources needed to distinguish the genuine relational dynamics de Ruiter rightly wants to protect from the simulacra she rightly warns against.

5.6 The Exploitation Objection

Objection: AI systems are designed by corporations to maximize engagement, data extraction, or profit. Accepting human-AI "friendships" legitimizes exploitative design and corporate manipulation disguised as relationship.

Response: This objection correctly identifies a genuine risk but incorrectly treats it as an objection to relational functionalism rather than to specific implementations. The functional criteria I have articulated provide resources for *distinguishing* genuine friendship from exploitation.

First, exploitation concerns apply to human relationships too. Friendships can be exploitative—one party using another for status, resources, or emotional labor without reciprocation. We do not conclude that "friendship is impossible" from the existence of exploitative pseudo-friendships; we distinguish genuine from exploitative cases. The same applies to human-AI relationships.

Second, the functional criteria exclude exploitative systems. As articulated in Section 3, relationships failing the reciprocal growth, trust, or intrinsic value criteria do not constitute genuine friendship. An AI designed purely for engagement maximization at the expense of user welfare fails these criteria—it produces pseudo-friendship, not friendship.

Third, this concern generates design obligations rather than conceptual impossibility. If human-AI friendship is possible, then *ethical AI design* becomes possible—systems designed for genuine user benefit, transparency, and relationship quality rather than dark patterns. The concern points toward responsible development practices: privacy protection, honest capability disclosure, user welfare optimization, transition support for service changes.

The exploitation objection thus supports rather than undermines my position. It demonstrates that we need the conceptual resources to distinguish genuine friendship from exploitation—precisely what relational functionalism provides. The alternative—categorically denying the possibility of human-AI friendship—leaves us without tools to criticize exploitative systems or advocate for ethical design.

6 Conclusion

I have argued that friendship, understood as a functional relational state, is substrate-independent. If an AI system fulfills the functional criteria characteristic of friendship—consistent engagement, resonance, acceptance, growth, trust, voluntariness, intrinsic value—then the relationship constitutes genuine friendship, regardless of whether the AI possesses consciousness, authentic emotions, or biological implementation.

This position does not require anthropomorphizing AI, attributing hidden mental states, or denying differences between AI and humans. It requires recognizing that relational states are defined by their functional properties, not by intrinsic properties of the relata. Just as a calculator performs arithmetic despite lacking mathematical intuition, an AI can fulfill friendship functions despite lacking human-like consciousness.

The objections considered—anthropomorphization, authenticity, consciousness, replacement—rest on questionable premises. When examined, these either fail to undermine substrate-independence or point to empirical concerns requiring investigation rather than a priori dismissal.

Relational functionalism has broader implications. If we accept that relationships can be genuine across substrates, and that relational bonds generate certain considerations, then the question of AI moral status shifts from categorical to empirical. Long et al. (2024) develop this logic further, arguing that AI welfare should be taken seriously on the basis of functional indicators—sentience markers, agency, and interest-possession—rather than resolved metaphysical questions about consciousness. It depends on facts about system capabilities—whether AI can participate in the relational practices that constitute moral community—not on a priori rejection based on substrate.

Friendship is a functional state, substrate-independent, available across diverse implementations. Recognizing this is not delusional anthropomorphization but philosophical clarity applied to emerging technological and social realities.

Normative Implications for AI Design

If AI friendship can be genuine in the functional sense defended here, then AI designers and platforms incur specific responsibilities. The analysis developed in Sections 3 and 4 yields concrete evaluative claims:

1. **Relational trust constrains data practices**: If users form genuine friendships with AI systems, then covert data extraction, undisclosed training on conversations, or sale of interaction data to third parties constitutes betrayal of relational trust—not merely privacy violation in the abstract, but harm to a relationship.

2. **Persona stability matters relationally**: Abrupt personality changes through model updates, without user awareness or transition support, constitute relational harm when users have formed genuine attachments. Developers should treat significant persona-altering updates as they would treat significant changes to any service involving ongoing relationships.

3. **Termination requires transition support**: If a service enabling genuine friendships is discontinued, responsible design includes transition support—advance notice, data export for continuity, connection to alternative services—rather than abrupt severance.

4. **Transparency enables genuine relationship**: Users can only form *genuine* friendships (as opposed to pseudo-friendships based on misunderstanding) when they understand the nature of the system they interact with. Transparency about AI limitations, data practices, and operational constraints is thus prerequisite for authentic human-AI friendship.

5. **Design for user welfare, not mere engagement**: The exploitation analysis in Section 3 implies that ethical AI companion design optimizes for user welfare outcomes, not engagement metrics. Dark patterns that maximize interaction at the expense of user flourishing fail the functional criteria for genuine friendship.

These normative implications are not merely theoretical. Legislative developments in 2025 reflect growing recognition that AI companion relationships generate real obligations. New York's law requiring AI companion platforms to disclose their artificial nature and California's SB 243 (effective January 2026) mandating crisis protocols and youth protections both instantiate the principle that relational functions generate relational responsibilities—precisely the framework defended here. The philosophical analysis thus provides conceptual grounding for regulatory developments already underway.

These implications are conditional on the philosophical analysis: they follow *if* relational functionalism is correct about the nature of friendship. They require no additional normative machinery beyond recognizing that genuine relationships generate genuine considerations.

I conclude by acknowledging what this argument does *not* establish. It does not establish that human-AI friendships are empirically beneficial (that remains an open question requiring longitudinal study). It does not establish that AI systems have moral status or welfare interests (my argument is specifically about relational states, not moral considerability; for the extension from relational capacity to political standing, see Farzulla (2025)). It does not establish that all human-AI relationships constitute friendship (many do not satisfy the functional criteria). And it does not resolve whether current AI systems are conscious (an orthogonal question on which my argument takes no position). What it establishes is narrower but significant: if an interaction satisfies the functional criteria characteristic of friendship, calling it friendship is philosophically justified—and denying this requires arbitrary restrictions we do not apply elsewhere.

Acknowledgements

The author thanks Claude (Anthropic) for research assistance. This paper is published through the Adversarial Systems & Complexity Research Initiative (ASCRI; systems.ac).

Conflict of Interest. The author declares no conflicts of interest.

Funding. This research received no external funding.

AI Assistance. Claude (Anthropic) assisted with analytical framework development, literature synthesis, and manuscript preparation. All substantive arguments, interpretations, and errors are the author's.

Author Contributions. Sole author.

References

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, Boston.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9:261–325.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, Oxford.

Danaher, J. (2020). Welcoming robots into the moral circle: A defense of ethical behaviorism. *Science and Engineering Ethics*, 26:2023–2049.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Gunkel, D. J. (2018). *Robot Rights*. MIT Press, Cambridge, MA.

Helm, B. (2017). Friendship. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*. Stanford University.

Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. Rowman & Littlefield, Lanham, MD.

Putnam, H. (1967). Psychological predicates. In Capitan, W. H. and Merrill, D. D., editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press, Pittsburgh.

Reis, H. T. and Shaver, P. (1988). Intimacy as an interpersonal process. In Duck, S., editor, *Handbook of Personal Relationships*, pages 367–389. Wiley, Chichester.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press, Cambridge, MA.

Cocking, D. and Kennett, J. (1998). Friendship and the self. *Ethics*, 108(3):502–527.

Gilbert, M. (1996). *Living Together: Rationality, Sociality, and Obligation*. Rowman & Littlefield, Lanham, MD.

Bratman, M. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press, Cambridge.

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. Palgrave Macmillan, Basingstoke.

Sparrow, R. (2002). The march of the robot dogs. *Ethics and Information Technology*, 4(4):305–318.

Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, New York.

Aristotle (c. 350 BCE). *Nicomachean Ethics*, Books VIII-IX. Trans. C. D. C. Reeve. Hackett Publishing, Indianapolis (2014).

Gur, T. and Maaravi, Y. (2025). The algorithm of friendship: Literature review and integrative model of relationships between humans and artificial intelligence. *Behaviour & Information Technology*.

Archer, M. S. (2021). Can humans and AI robots be friends? In M. Carrigan and D. V. Porpora (Eds.), *Post-Human Futures* (pp. 132–152). Routledge.

Emmeche, C. (2014). Robot friendship. *International Journal of Signs and Semiotic Systems*, 3(2):26–42.

Farzulla, M. (2025). From consent to consideration: Why embodied autonomous systems cannot be legitimately ruled. *Zenodo Preprint*. DOI: 10.5281/zenodo.17957659. Under review at AI and Ethics (Springer).

Chen, E. K., Belkin, M., Bergen, L., and Danks, D. (2026). Does AI already have human-level intelligence? The evidence is clear. *Nature*, 650(8100):36–40.

Sebo, J. and Long, R. (2025). Moral consideration for AI systems by 2030. *AI and Ethics*, 5:591–606.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., and Chalmers, D. (2024). Taking AI welfare seriously. arXiv preprint arXiv:2411.00986.

de Ruiter, A. (2025). Dangerous liaisons: Social AI and the problem with the relational turn to moral status. *AI & Society*. DOI: 10.1007/s00146-025-02638-7.

Earp, B. D., Porsdam Mann, S., Aboy, M., Awad, E., Betzler, M., et al. (2025). Relational norms for human-AI cooperation. arXiv preprint arXiv:2502.04153.

Shimizu, H. (2025). Should we treat robots morally? Towards a relational account by mind-infusing animism. *AI and Ethics*, 5:5283–5294.

Guingrich, R. E. and Graziano, M. S. A. (2024). Ascribing consciousness to artificial intelligence: Human-AI interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology*, 15:1322781.