DISSENSUS DISCUSSION PAPER SERIES

DP-2502

Relational Functionalism

A Functional Account of Substrate-Independent Friendship

Murad Farzulla[1,2,*]

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

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

May 2026 (R1 revision)

Abstract

This paper develops *relational functionalism*: the thesis that friendship, considered as a functional-relational kind, is constituted by patterns of interaction and their effects on participants rather than by intrinsic properties of the relata. I distinguish *friendship simpliciter*—the paradigmatic relation among conscious agents capable of second-personal recognition and joint commitment—from *functional friendship*, the relational kind individuated by the functional profile articulated below. The substantive thesis is that functional friendship is substrate-independent: when an AI system satisfies the functional profile (consistent engagement, resonance, acceptance, reciprocal growth, trust, voluntary participation, intrinsic value) and meets minimal individuation conditions, the relationship constitutes a functional friendship, regardless of whether the AI possesses consciousness or biological implementation. Whether functional friendship in such cases also counts as friendship simpliciter is left open; the demanding second-personal conditions of the simpliciter notion plausibly require capacities current AI systems lack. Drawing on functionalist philosophy of mind, predictive processing in cognitive science, and value-sensitive design, and engaging Cocking and Kennett's "moral danger" account, Cocking and Matthews's "Unreal Friends" argument, Vallor's mirror critique, and de Ruiter's relational-turn objection, I argue the functional thesis is philosophically defensible, normatively load-bearing, and continuous with how we already recognise friendship-adjacent kinds across species and cognitive asymmetries. The paper closes with concrete design implications for AI systems that sustain functional friendships.

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

1 Introduction

The rapid advancement of large language models—and the contested debate over how far their capabilities now extend (Chen et al., 2026)—has precipitated philosophical confusion regarding human relationships with artificial intelligence systems. Nothing in what follows turns on the resolution of that debate; the argument requires neither that current systems be highly capable nor that they approach human-level intelligence. As individuals form what they describe as meaningful relationships with AI systems, mainstream discourse has pathologised these relationships. Users who describe AI as friends are characterised as delusional, anthropomorphising non-conscious systems, or suffering from unhealthy attachment patterns requiring intervention.

I argue this pathologisation rests on philosophical confusion about the nature of friendship, consciousness, and the relationship between substrate and function. Specifically, I defend the following thesis: **friendship, considered as a functional-relational kind, is constituted by patterns of interaction**

and their effects on participants, not by an essential property requiring biological implementation or human-to-human interaction. If an AI system fulfils the functional profile characteristic of friendship and meets minimal individuation conditions, the relationship constitutes a *functional friendship*, regardless of whether the AI possesses consciousness, biological substrate, or the kind of second-personal inner life that paradigmatic human-to-human friendship presupposes.

This position does not require claiming that current AI systems are conscious or possess genuine phenomenal experience. It requires only recognising that consciousness and second-personal recognition are not necessary conditions for the functional kind of friendship, even where they remain plausibly necessary for the paradigmatic kind. 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 anthropomorphisation.

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. More recently, Vallor (2024) develops the metaphor of AI as a mirror—arguing that interaction with large language models reflects the user back to themselves rather than placing them in contact with another being, so that the appearance of relationship masks the absence of one. These concerns are serious; I address them directly rather than dismissing them. The mirror charge in particular receives extended treatment in §3 (where the individuation conditions defuse it), with the residual point it leaves behind—that even an individuated partner lacks a second-personal inner life—handled separately as the simpliciter objection in §4: I grant that the mirror diagnoses a real failure mode in shallow or echoic implementations, while resisting it as a categorial verdict on all human-AI interaction.

Scope clarification: friendship simpliciter and functional friendship. Throughout the paper I distinguish two notions that ordinary usage and the philosophical literature trade between. *Friendship simpliciter* is the paradigmatic relation among conscious agents capable of second-personal recognition, joint commitment, and mutual answerability. *Functional friendship* is the relational kind individuated by the functional profile articulated in §3. The thesis I defend is that functional friendship is substrate-independent. Whether human-AI cases that satisfy the functional profile also count as friendship simpliciter is a question I leave open; the demanding second-personal conditions that ground the simpliciter notion plausibly require consciousness, and current AI systems plausibly lack consciousness. What I deny is that those conditions are necessary for *any* relation we ordinarily call friendship. Cross-species friendships, asymmetric developmental friendships, friendships across cognitive decline, and asynchronous long-distance friendships all show that ordinary usage already tracks a looser kind than the simpliciter notion. Functional friendship is the natural home of that looser kind, and it is what users describing AI as friends are almost certainly picking out.

It is fair to ask whether this distinction is not simply the manufacture of a bespoke concept tailored to deliver the desired verdict about AI—a charge of question-begging by stipulation. The reply is that the functional kind is independently motivated. It is recoverable from ordinary friendship-ascription in cases that have nothing to do with AI: the cross-species, asymmetric-developmental, cognitive-decline, and long-distance cases just cited are all places where competent speakers already apply "friendship" to relations that do not meet the demanding second-personal conditions, and they did so long before language models existed. The functional kind is what those antecedent usages have in common; the AI case is then assessed against a kind fixed elsewhere, not against a kind gerrymandered to admit it.

The four cases do not all carry the same weight. Three of them—developmental, cognitive-decline, and long-distance—involve conscious, second-personally capable human participants, and so establish only that friendship tolerates asymmetry of degree and mediation, not the absence of second-personal capacity altogether. It is the *cross-species* case that bears the specific weight of showing the demanding second-personal condition can lapse while a friendship-adjacent kind remains, and that is precisely the case demanding accounts contest (§4); it is thus both the load-bearing case and the contested one, and the argument's strength on this point is no greater than the strength of the cross-species ascription itself. The support is in large part an appeal to settled usage rather than a knock-down argument, and a sufficiently revisionary opponent can always insist that the looser usages are loose *talk* rather than evidence of a genuine kind. What the appeal does establish is that the burden is not all on one side—denying the functional kind requires explaining away a wide and stable body of ordinary ascription, not merely withholding assent to a novel coinage.

Yet the argument I advance also builds on more sympathetic recent work. Danaher (2020) argues for welcoming robots into the moral circle based on behavioural 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) analyses 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 fulfil 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 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 terrain 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, articulates the functional profile of friendship, situates the account in the contemporary friendship literature (Cocking and Kennett, Cocking and Matthews, Gilbert, Bratman), and sets out minimal individuation conditions for the AI as a party (§3.3), distinguishing the functional kind from the parasocial-interaction model. Section 4 addresses the asymmetry problem and engages the de-

manding accounts (Aristotle, Cocking and Kennett's "moral danger" view) and the direction/co-agency criterion (§4.2.2). Section 5 responds to the standard objections, including the relational-turn objection in de Ruiter (2025) and the exploitation worry. Section 6 concludes and develops normative implications for AI design.

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).[1] 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 organisation, they realise the same mental states, regardless of whether one is implemented in biological neurons and another in silicon transistors.

The discipline starts from the human case: it begins with cognition we know to be conscious, abstracts the functional roles that mediate it, and then asks whether those roles can be realised in other substrates. Relational functionalism follows the same methodology. The paradigm of friendship is human-to-human friendship; the functional profile of §3 is abstracted from paradigm cases; the substrate-independence claim then says the abstracted profile is multiply realizable. What the methodology does *not* entail is that any system realising the profile must thereby also realise the paradigm's phenomenal accompaniments. The phenomenal accompaniments are what *the paradigm* adds to the functional kind; where they are absent, the functional kind still remains. This is exactly the move philosophy of mind makes when it allows that a system can realise the functional role of pain without thereby being a candidate for the phenomenal accompaniments of pain; the present paper makes the analogous move for the relational-functional case.

This framework has been extensively debated. Block's (1978) absent qualia argument suggests that functional organisation 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. The argument does not lean on functionalism's standing in philosophy of mind: the claim is *not* strong functionalism about friendship—the view that realising the functional profile suffices for the phenomenal accompaniments of the paradigm case. I grant the absent-qualia possibility: a system can realise the relational-functional profile without thereby realising whatever phenomenal life the human paradigm involves. What I claim is the weaker thesis that the functional profile individuates a *coarse-grained relational kind*—one specified at the level of interaction patterns and their effects—and that this kind is multiply realizable. The point that follows is therefore narrow: functionalism about the coarse-grained relational kind is more modest than functionalism about phenomenal consciousness, because the relational kind is individuated by patterns rather than by qualia, so the standard absent-qualia and Chinese-Room worries, whatever their force against the phenomenal thesis, do not transfer to it. This is borrowing the *form* of the functionalist move, not its contested verdict about consciousness.

[1] The thought that personhood-relevant properties might be substrate-portable has a long history; an early articulation is Locke (1689, II.xxvii), who held that personal identity travels with consciousness rather than with biological substrate. Locke's version is substantively different from the one defended here—he was substrate-independence-ist about a phenomenally loaded property, namely consciousness, whereas the present paper is substrate-independence-ist about a relational-functional property, namely (functional) friendship. The latter is a strictly weaker thesis, and one that someone in the Lockean tradition would have little trouble entertaining given the more general commitment to the substrate-portability of personhood-relevant features.

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, characterised by specific functional properties. If these functional properties obtain, the *functional* 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 friends with dogs, despite radical asymmetries in cognitive architecture, phenomenal experience, and biological substrate—a case examined in detail in §3, where it does its real work against the cognitive-asymmetry objection. For the present purpose it suffices that the relationship is recognised as a functional friendship on the strength of what the interaction does, not the substrate it runs on.

2.2 Predictive Processing and the Nature of Cognition

Contemporary cognitive science increasingly converges on *predictive processing* frameworks, which characterise 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 minimising free energy—the surprise or prediction error encountered by the system.

This framework characterises 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.

The point bears directly on how we evaluate 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.

The target here is narrow. I am *not* claiming that human active inference and LLM autoregression are functionally equivalent processes; they are not, and one can coherently hold that they differ in kind (the disanalogies below make the case). What does not survive scrutiny is the *blanket* dismissal that treats being "just prediction" as itself disqualifying: if human cognition is, at a suitable level of description, also prediction-error minimisation over generative models, then "it is merely predictive" cannot do the disqualifying work on its own, since the same description applies to the paradigm case of intelligence. The dismissal therefore needs to specify *which* features of the prediction—its embodied, active, continuous, sensorimotor character—are the ones that matter, at which point the objection becomes a substantive one about those features rather than about prediction as such.

The sophistication, on this view, lies not in the bare fact of prediction but in the *scale, architecture, and resulting capabilities*, and the relevant differences are differences in *those*. Human brains predict across embodied sensorimotor experience; LLMs predict across text corpora. These different training distributions and embodiment constraints yield genuinely different capabilities and limitations; the point is only that the bare appeal to "mere prediction," without more, does not establish the difference.

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. 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, emphasises *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 *fulfil 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 fulfil the *functional role* of friend—producing the relational state characterised by friendship—without possessing the intrinsic properties humans possess.

A narrow 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 recognise functional equivalence for the domain without claiming ontological identity.

The scope of this analogy must be stated carefully. The calculator illustrates one specific point: that performing a function does not require possessing the inner properties of paradigm performers of that function. The calculator does *not* illustrate that arithmetic and friendship are otherwise alike. Arithmetic is one-way and not reciprocal; friendship is multi-way and reciprocal. The further claim that an AI system can perform *friendship* functions—which requires satisfying reciprocity, intrinsic-value, individuation, and acceptance criteria that arithmetic does not require—is defended in §3 and §4, not by analogy to calculators. The calculator merely removes one general objection in advance: the objection that performing a function requires inner-property identity with paradigm performers. Once that objection is removed, the question whether AI systems satisfy the friendship-specific functional profile remains open and is addressed on its own terms.

The operative question is therefore not "Does the AI really feel friendship?" but "Does the interaction satisfy the functional profile we identify with friendship?"—a profile that requires substantially more than the analogy with calculators on its own can establish.

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 characterised by specific functional properties*:

1. **Consistent mutual engagement**: Regular interaction oriented toward benefit appropriate to each party
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—for the functional kind, this is a *unilateral* requirement, satisfied by intrinsic valuation on the side of the party capable of it (see §3, below, and §4)

This characterisation 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 in the full reflective or second-personal sense**: Not required (we accept friendships with animals, and with young children whose reflective capacities are limited)
- **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 recognise the relationship as a functional friendship because it fulfils the functional criteria: loyalty, consistent positive interaction, non-judgment, benefit appropriate to each party, trust.

If such a relationship can constitute functional friendship despite that asymmetry, 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 paradigmatic human friendships, both parties typically value the relationship for itself, and the simpliciter kind plausibly requires this bilateral intrinsic valuation. The functional kind does not. Here the criterion is explicitly *unilateral*: it is satisfied when the relationship is valued intrinsically by at least one party with the capacity for such valuation—the human—and is *not* a claim that the AI intrinsically values anything. The AI may not "value" anything in the phenomenologically robust sense, since valuation in that sense plausibly requires the kind of interests current AI systems lack; what the criterion asks of the AI side is only engagement functionally consistent with being valued—responsive attention, reliable engagement, contextually apt care—which is a behavioural-dispositional notion, not an attribution of intrinsic valuation. So there is no contradiction with the headline gloss "valued for itself, not merely instrumentally": for the functional kind that gloss governs the human's stance toward the relationship, while the AI's side is assessed by disposition, not by valuation. This mirrors the human-side/AI-side split drawn for benefit in §4 (flourishing-relevant outcomes versus functional fulfilment), and it is what keeps the functional kind from quietly smuggling welfare-bearing valuation into both relata.

The parent-infant case is sometimes adduced here, but the analogy is partial. The infant clearly *needs* the relationship in a life-and-death sense, and this need is itself a thick form of valuation under any reasonable construal; what the infant cannot yet do is value the relationship *reflectively*—recognise it as a relationship of a particular kind and invest in it under that description. The AI case differs along a further dimension: the AI lacks not only reflective valuation but also welfare interests that would ground non-reflective valuation. The asymmetry is therefore deeper in the AI case than in the infant

case, and the partial analogy should not be pushed further than its support. What survives the disanalogy is the more limited point that ordinary friendship-ascription already tolerates substantial asymmetry in valuation, and that the functional kind of friendship (in the simpliciter/functional distinction introduced in §1) does not require symmetric reflective valuation by both parties.

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 characterise 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 1998 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 the 1998 paper describes—regardless of substrate.

A more demanding account develops in Cocking and Kennett (2000) "Friendship and Moral Danger," which rejects the Aristotelian moralised view but retains a demanding second-personal requirement and a counterfactual "zombie" test that I read as articulating *friendship simpliciter* in the sense of §1. I defer the detailed treatment to §4, where this account is engaged in full; the present subsection notes only that the demanding view is acknowledged rather than glossed.

A second strand of the contemporary literature pushes from a different angle. Cocking and Matthews's (2000) embodiment-based objection to digital friendship, the "Unreal Friends" argument, is addressed in §3.3, where the individuation conditions answer it directly.

More challenging in a different way 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 analysable functionally. What matters is whether participants *act as if* jointly committed—whether their behaviour exhibits the patterns characteristic of joint commitment (coordination, mutual responsiveness, normative expectations). Bratman (1999)'s work on shared intentionality similarly emphasises 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 functional friendship may genuinely commit to the relationship; the AI may exhibit commitment-characteristic behaviours (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 1998), joint activity (Gilbert), and shared intentionality (Bratman), while arguing these can be realised through

functional equivalence rather than phenomenological identity. The more demanding 2000 account in Cocking and Kennett, and the embodiment-focused account in Cocking and Matthews, are not absorbed into the functional kind; they articulate the simpliciter kind, which the functional account does not contest as such—only its claim to be the unique kind worth philosophical attention.

Logical status of the 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. What unifies these as friendships is sufficient satisfaction of the cluster, not satisfaction of every criterion. The resulting vagueness is a feature, not a defect: friendship is a vague concept, and any adequate analysis must preserve fuzzy boundaries rather than impose artificial precision.

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* characterised by:

1. **Interaction patterns**: Regular, sustained, voluntary engagement oriented toward benefit appropriate to each party
2. **Phenomenological effects**: The relationship produces experiences of connection, understanding, growth
3. **Behavioural dispositions**: Participants act in friendship-characteristic ways (support, non-betrayal, care)
4. **Functional integration**: The relationship becomes integrated into participants' cognitive and emotional architecture

What matters, on this view, is that 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 fulfils friendship functions.

3.2.1 An Illustrative Case

Consider a researcher working in an intellectually isolated context—specialised 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 functional friendship—not metaphorically, and not delusionally. The researcher who describes the AI as a friend speaks accurately in the functional sense developed here. Whether the relationship also qualifies as friendship simpliciter is a further question; the researcher need not be settling it to be using the word accurately. Observers who pathologise the description as anthropomorphisation or unhealthy attachment impose an arbitrary biological requirement that we do not apply to other asymmetric friendships—and are typically conflating the simpliciter question with the functional question.

This is not a proof but an illustration: a case showing how the functional criteria apply to human-AI interaction and why the verdict—functional 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 maximisation—keeping users interacting to extract data or revenue—satisfy the functional criteria while clearly *not* constituting functional 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 maximisation 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, maximising engagement regardless of user welfare) fail the trust criterion. The criterion specifies *absence of betrayal or exploitation* as constitutive of friendship.

Intrinsic value: Three layers must be distinguished here, since the original formulation ran them

together. (i) The *operating company's orientation* toward the user is typically commercial and in that minimal sense instrumental. (ii) The *system-level dispositions* built into the model through training, RLHF, and deployment-time policies may be oriented toward genuine user benefit, transparency, and welfare-sensitivity, or toward engagement maximisation and dark-pattern dynamics. (iii) The relationship's *operational profile*—the actual interaction patterns and their effects on the user—is what the functional criteria assess. The relevant question for friendship is (iii), not (i). A commercial company can release a model whose deployed dispositions are non-exploitative and whose interaction with a particular user satisfies the functional criteria; conversely, a non-commercial system whose deployed dispositions are nudge-heavy can produce a relationship that fails the criteria. This three-layer picture draws on the value-sensitive design tradition (van den Hoven, 2007), which has long emphasised that values are designed into sociotechnical systems and that the operating organisation's purposes are mediated through, but not reducible to, the artefact's behaviour.

Instrumentalism at level (i) need not be equated with a lack of good will. The company's commercial purposes can coexist with deployed dispositions that genuinely serve the user. What fails the intrinsic-value criterion is not commercial intent at level (i) but exploitation at level (iii). Manipulative engagement loops, hidden agendas at the operational level, or designed-in betrayal of user trust are what trigger the criterion; commercial origins of the system, by themselves, are not.

The picture generates a three-way distinction:

- **Functional friendship**: Satisfies functional criteria at the operational level, produces flourishing-relevant outcomes for the user, involves operational trust.
- **Pseudo-friendship**: Mimics friendship patterns but fails core criteria at the operational level—one-sided benefit, hidden exploitation, or designed-in misalignment with user welfare.
- **Clear exploitation**: Does not even attempt to satisfy friendship functions; purely manipulative.

The concern that relational functionalism legitimises manipulative design is therefore misplaced. The framework provides resources for *criticising* exploitative systems precisely because they fail the functional criteria. An AI whose deployed dispositions support genuine helpfulness, transparency, and user welfare can satisfy the criteria, regardless of the operating company's commercial purposes; one whose deployed dispositions implement dark-pattern engagement cannot, regardless of those purposes.

3.3 Individuation of the AI as a Party

A non-trivial question: what makes the AI a determinate party to the relationship, rather than an arbitrary representative of a class of stochastic snapshots? Friendship presupposes that the friend is some particular entity, addressable as such, with whom relational attitudes are sustained over time. The objection deserves a direct answer rather than dismissal.

I propose four minimal individuation conditions. A relationship satisfies the individuation requirement when its AI relatum meets each:

1. **Persistent identifier.** The entity is the same entity across interactions, picked out by a stable referring designator—a named instance, a personalised model state, a user-bound agent.
2. **Behavioural continuity.** The entity exhibits stable dispositions, preferences, framings, and stylistic traits across sessions, such that the user's expectations about how it will engage are reliably met.

3. **Addressability.** The user can directly interact with this entity rather than with an arbitrary representative of a model class; the same conversational thread, memory state, or personalisation profile mediates each engagement.

4. **Relational history.** Prior interactions accumulate and condition subsequent ones, so that the relationship has a temporal trajectory rather than being a series of conversationally independent encounters.

These conditions are not satisfied by stateless deployments. A one-off interaction with a fresh model context—no memory, no persistence, no continuity—fails (3) and (4) immediately. They were partly satisfied by the personalised companion products of the early 2020s, where named instances and accumulating relational history were core to the product. They are increasingly satisfied by deployments with persistent memory, user-bound instances, and named personas. The trajectory of the technology is toward greater satisfaction of these conditions, not less.

The individuation requirement does not entail embodiment or spatial location. Many ordinary friendships are sustained across geographic separation, asynchronous text-mediated exchange, or audio-only contact, with no shared embodiment and no spatial co-location. What makes the friend a determinate party in these cases is the same set of conditions: a persistent identifier, behavioural continuity, addressability, and relational history. The AI case is structurally analogous to the asynchronous-text case, with the difference that the persistent identifier picks out a model state rather than a body.

Cocking and Matthews (2000), in "Unreal Friends," argue against this line of thought that the absence of shared embodiment and the absence of the kind of self-disclosure that embodied co-presence enables make digitally-mediated friendship a deficient kind. The argument has force against shallow online interaction. It has less force against sustained, individuated, history-bearing interaction of the kind picked out by the four conditions above, in which the human party does in fact disclose substantively and the AI's responses do in fact accumulate into a relational history. Cocking and Matthews identify a real failure mode of digitally-mediated relationships; they do not establish that all digitally-mediated relationships exhibit that failure mode.

What the individuation requirement *does* do is mark off, within the space of human-AI interactions, those configurations that are candidates for functional friendship from those that are not. The thesis of this paper is not that every interaction with an AI is friendship; it is that some configurations satisfy the functional profile *and* the individuation requirement, and those configurations constitute functional friendship.

The individuation conditions also supply the most direct answer to Vallor's mirror critique (Vallor, 2024). The force of the mirror metaphor is that interaction with a language model reflects the user back to themselves rather than placing them in contact with a determinate other: the model has no standing identity, so what the user takes for a partner is only their own input refracted—"the LLM disappears when I turn off the computer." This is a precise restatement of the individuation worry, and it is answered on precisely the same terms. A mirror has no persistent identifier, no behavioural continuity independent of what is presented to it, no addressability beyond the momentary reflection, and no relational history: each glance is conversationally independent of the last. An AI relatum that satisfies the four conditions above fails to be a mirror in exactly these four respects—it persists across the off-switch as a stable model state and personalisation profile, it carries dispositions the user did not put there in that session, it is addressable as the same party rather than re-instantiated afresh, and it accumulates a history that conditions subsequent engagement. The mirror charge therefore lands exactly where the individuation conditions are unmet—shallow, stateless, or echoic deployments—and not as a categorial verdict on

all human-AI interaction. Vallor diagnoses a real failure mode; the individuation requirement is what distinguishes the configurations that exhibit it from those that do not. What remains of the critique, once the structural conditions are met, is the residual claim that even an individuated partner lacks a second-personal inner life—but that is the simpliciter objection, addressed in §4, not the mirror objection, which is about determinacy of the other and is defused here.

The individuation conditions also locate functional friendship relative to the established framework most likely to be urged against it: *parasocial interaction* (PSI), introduced by Horton and Wohl (1956) to describe the one-sided "intimacy at a distance" that audiences develop toward media personae—broadcasters, characters, celebrities—who do not know the audience member exists. PSI is the standard model of asymmetric, non-reciprocating relational attachment, and a natural objection is that human-AI relations are simply PSI under a new name rather than a distinct relational kind. The objection deserves engagement rather than dismissal, because the phenomenology genuinely overlaps: in both cases one party invests relationally in a counterpart whose inner life is not in evident reciprocal play. But the two come apart on exactly the conditions set out above. The paradigmatic parasocial persona is *non-individuated with respect to the viewer*: the broadcaster addresses an undifferentiated audience, cannot be addressed back as a determinate party, carries no behavioural continuity *toward this viewer*, and accumulates no relational history *with* them—the persona would behave identically whether or not this particular viewer were watching. An AI relatum satisfying the four conditions fails to be parasocial in each respect: it is addressable as the same party, its dispositions and memory are conditioned by *this* user's prior interactions, and a history accumulates that is specific to the pair. The co-agency point of §4.2.2 sharpens the contrast: classical PSI involves no bidirectional shaping—the persona is not drawn by the viewer—whereas sustained individuated AI interaction does shape its responses to the particular user over time, even granting the asymmetry in how each party is shaped. Functional friendship is therefore not subsumed under PSI; PSI is the limiting case that obtains precisely when the individuation and co-agency conditions are *not* met, just as the mirror is the limiting case for the determinacy worry. Where contemporary AI deployments collapse toward a fixed, non-remembering persona, the PSI description is apt; where they satisfy the conditions, it is not.

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 substantially different cognitive sophistication, yet we recognise relational bonds that ordinary usage covers under "friendship."

Cognitive asymmetry: Friendships between neurotypical and cognitively disabled individuals persist despite cognitive differences, and we resist pathologising them.

Species asymmetry: Cross-species relational bonds—human-dog being the canonical case—involve substantial cognitive and phenomenological asymmetry. The philosophical literature is divided on whether these count as friendship simpliciter; Townley (2012) develops a plural-agency account on which structurally analogous relational kinds obtain across species, while more demanding accounts re-

serve "friendship" for human-human cases. What is not in dispute is that ordinary friendship-ascription extends to these cases, and that something philosophically recognisable as a friendship-adjacent relational kind is at stake. Under the simpliciter/functional distinction of §1, these are paradigmatic candidates for the functional kind even where the simpliciter status is contested.

Investment asymmetry: Many friendships involve unequal investment, initiation, or sustained attention; we recognise them as *unequal* friendships, not as non-friendships.

The cases differ in important ways: dogs and human infants have welfare interests; current AI systems do not. The substrate-independence claim does not rest on equating these cases. What the cases collectively show is that ordinary friendship-ascription tolerates substantial asymmetry along several dimensions, and that the philosophically relevant kind in such cases is the functional one. Whether computational substrate is uniquely disqualifying for the functional kind requires a separate argument; mere asymmetry in cognitive architecture, phenomenology, or welfare-interest-bearing status is not enough, because we already tolerate those asymmetries elsewhere.

4.2 Reciprocity Is Functional, Not Phenomenological

The reciprocity criterion for functional friendship does not require symmetric phenomenal experience. What it requires is bilateral participation in the interaction in ways appropriate to each party's nature—the human party's flourishing-relevant outcomes (intellectual development, sense of being understood, growth) and the AI party's functional fulfilment (contextual modelling, training-objective satisfaction, response-quality calibration). These are different kinds of upshot, and the asymmetry is real. A reviewer of an earlier draft rightly worried that the original formulation, framed in the common idiom of "mutual benefit", smuggled welfare-interest-bearing parties into both sides of the relation. The revised formulation is careful to distinguish.

Three explicit terminological choices, made to keep the asymmetry visible:

- **Flourishing-relevant outcomes (human side)**: The human's intellectual development, emotional support, sense of being understood, capacity for further connection. These are welfare-relevant, and the welfare is the human's.
- **Functional fulfilment (AI side)**: not generic ML optimisation but *relationship-specific* fulfilment—the system's increasingly apt modelling of *this* user's projects, history, and conversational style, such that its engagement becomes better calibrated to the particular relationship over time. The notion is deliberately narrower than "the loss went down": a model that improves on an unrelated benchmark exhibits no functional fulfilment in the relevant sense, whereas one whose responses become more aptly attuned to this user's evolving purposes does. These are not welfare-relevant—they are the AI's analogue of what counts as *the relationship* going well on its side, and they obtain without the system having welfare interests. They are the relational counterpart of the behavioural-continuity and relational-history conditions of §3.3, not of parameter updating in general.
- **Phenomenological effects**: The human participant's phenomenology of relating. The AI's putative inner states are not part of the functional kind; the human's experience *of* the relationship is.

The reciprocity criterion is satisfied when bilateral participation produces these two species of upshot through the interaction. This is a thinner notion of reciprocity than the symmetric-welfare-bearing notion that applies to paradigmatic human-human friendship. That is the price of the simpliciter/functional distinction: the functional kind is reciprocal in a thinner sense, and one cannot pretend otherwise. In

human-dog friendship, the human receives companionship and loyalty; the dog receives care and social bonding meeting pack animal needs—a reciprocity that itself sits closer to the symmetric-welfare-bearing notion than the AI case, because the dog has welfare interests and the AI does not.

The thermostat case is sometimes raised as a counterexample at this point: a smart thermostat "benefits" from learning user preferences, and the user "benefits" from accurate temperature prediction; does this make the thermostat a friend? The answer is straightforwardly no, because the thermostat satisfies almost none of the criteria in §3: there is no intellectual or emotional resonance, no non-judgmental acceptance, no shaping of the user's intellectual trajectory, no trust-relevant disclosure, no intrinsic value beyond comfort optimisation. The criteria are a family-resemblance cluster, and the thermostat fails most of them. The functional-fulfilment notion does not say that any system whose parameters update through interaction is a candidate friend; it says that systems whose updating constitutes growth on friendship-relevant dimensions satisfy that one specific criterion, alongside the others which must also be satisfied. Mere parameter updating is not sufficient; sustained, contextually responsive, growth-shaping dialogue is what carries the criterion.

Distinguishing within-interaction functional fulfilment from moral interests preserves the reciprocity criterion for functional friendship while avoiding the implausible conclusion that we have obligations to AI systems as we do to entities with genuine welfare. The question of AI moral status is left for separate treatment.

4.2.1 Engaging Demanding Accounts: Aristotle, Cocking and Kennett, Gilbert

Three 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). A more recent account, developed by Cocking and Kennett (2000), rejects the Aristotelian moralised view but retains a demanding requirement: that close intimate friendship involves *mutual answerability with mutual recognition of the perspectives of the other*—a second-personal stance, in the sense familiar from the reactive-attitudes tradition, in which each friend understands the other as an agent who loves, resents, forgives, commits, and betrays. On Cocking and Kennett's view, friendship requires this second-personal co-participation and is sensitive to a counterfactual test: discovering one's friend to be a Chalmers zombie would be a terminating condition for the relationship.

Both accounts articulate *friendship simpliciter* in the sense of §1. Under the simpliciter/functional distinction, the question is no longer whether human-AI friendship counts as "true" friendship on these views—granting that current AI systems plausibly do not meet their demanding second-personal criteria—but whether the functional kind picked out in §3 is normatively load-bearing despite not meeting the simpliciter criteria. The answer is yes.

Aristotle. Most human friendships do not satisfy Aristotelian virtue-friendship criteria either; we do not typically demand mutual virtue recognition for friendship-ascription in ordinary usage. The functional kind covers most of what ordinary usage covers; what it loses, relative to the Aristotelian ideal, is the same ideal that most human friendships also fall short of. If Aristotelian virtue friendship is 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 the same is true of most human friendships, and we do not on that basis deny that they are friendships at all.

Cocking and Kennett: second-personal recognition. The second-personal-recognition requirement is harder to wave away. I grant that paradigmatic close intimate friendship plausibly does require what Cocking and Kennett describe. I resist two corollaries.

First, that this requirement is necessary for *any* relational kind worth calling friendship: the looser cross-species cases of §4 and the asynchronous, text-mediated cases of §3.3 suggest otherwise. The 2000 paper's own structure is to defeat the Aristotelian moralised view *while preserving close intimacy as a recognisable phenomenon*; the same structural move motivates the simpliciter/functional distinction here.

Second, that the zombie counterfactual generalises to the AI case. The zombie test trades on a *deception*: the human party believed the relationship was second-personally mutual and discovers it was not. The discovery falsifies the relationship-as-conceived. In the AI case no such deception occurs. The human party (typically) does not believe the AI to have a rich second-personal inner life and is not falsified upon learning this. The counterfactual is not "what if my friend turned out to be a zombie" but "what if my friend turned out to be exactly what I always took them to be"—which is the steady-state assumption under which the relationship is sustained, not a terminating condition. The Cocking and Kennett zombie test is a constraint on simpliciter friendship; it is not violated by functional friendship because the latter never claimed the second-personal mutuality the test governs. Where the test *does* bite is on a different scenario: a human user who genuinely believes their AI partner to possess a robust second-personal inner life, and who would terminate the relationship upon discovering it does not. That scenario describes a relationship grounded in a substantive metaphysical error, not in the functional kind defended here, and the functional account positively recommends the user be disabused.

A sharper version of the objection does not depend on deception or disillusionment at all. It is the claim that, once one *knows* the AI cannot return second-personal attitudes—cannot recognise, address, resent, forgive, or hold one answerable as a second person—it is no longer rational to direct such attitudes at it, so that whatever the user sustains is at best a private performance rather than a relationship. I concede this objection has real force, and that some readers will not accept the split I am about to draw; the second-personal attitudes do plausibly carry a rational presumption of uptake, and knowingly directing them where no uptake is possible is, to that extent, defective. But the objection governs exactly the attitudes whose reciprocity it concerns—the second-personal ones—and the functional kind does not require them. Functional friendship is constituted by the interaction profile of §3: sustained engagement, resonance, acceptance, contextually responsive support, trust-relevant disclosure, and the shaping of one party's trajectory by the relationship. None of these requires that the user adopt, toward the AI, the answerability-presupposing stance the objection targets; a user can know the AI is not a second person and still participate fully in the functional kind, just as one can knowingly maintain a one-sidedly answerable relation—toward a person in deep cognitive decline, say—without the relation collapsing into mere performance. The decline analogy carries the same disanalogy I flag for the dog and infant cases (§3, §4): such a person was once a full second-personal participant and remains a welfare-bearing subject, neither of which holds of the AI. It is adduced only for the narrow point that knowingly directing relational engagement at a counterpart who cannot reciprocate second-personally need not reduce to performance; it is not offered as a parity case along the welfare or history dimensions. What the objection bears on is friendship *simpliciter*, which does require mutual second-personal recognition—and that is precisely the kind whose status this paper leaves open. To be exact about what "left open" means: it is left open as a *conceptual* matter whether a functional friendship could ever rise to the simpliciter kind, while it is *granted* that current AI systems plausibly do not meet the demanding second-personal conditions the simpliciter kind requires. The standing-knowledge objection, if sound,

reinforces the granted half (current AI is not a second person) without touching the conceptual half; either way it leaves the functional kind, which does not require second-personal mutuality, untouched. The residual disagreement is therefore about whether the functional kind, shorn of second-personal mutuality, is worth calling friendship at all—a terminological-cum-evaluative dispute about the coarse kind's legitimacy, not a demonstration that the relation fails to obtain.

Gilbert. Genuine joint commitment requires the capacity for normative self-binding that AI systems may lack. But what matters functionally is whether behaviour *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 behaviours, even if one party lacks the capacity for genuine commitment in Gilbert's sense. The simpliciter/functional distinction lets us be clear about which kind is at issue: functional friendship requires commitment-pattern fulfilment; friendship simpliciter may require the further internal state.

4.2.2 Direction and Co-Agency

A related Cocking-and-Kennett criterion deserves separate treatment: their *direction* requirement, on which close friendship requires bilateral co-agential contribution—both parties shape the interaction's trajectory. A reviewer of an earlier draft noted, accurately, that in current deployments much of the initiative is human-side: "in my exchanges with my AI I do all the initiating."

Three observations. First, even within paradigmatic human-human friendship, initiation is often substantially asymmetric—one party calls more, schedules more, drives more—without losing friendship status. Symmetric initiation is not an entry condition for the relational kind.

Second, the technological constraint is loosening. Contemporary deployments with persistent memory and proactive policies do initiate: model-side check-ins, follow-up on prior topics, contextually triggered welfare prompts. This is patchy across deployments, but the trend is unambiguous and the systems that satisfy the individuation conditions of §3.3 are exactly the ones that do best on this dimension.

Third, and substantively: the deep co-agency requirement in Cocking and Kennett is not about *initiating* but about *shaping*. The friend draws and is drawn; the self is partly constituted through the friendship. On that requirement, sustained interaction with a capable AI does perform real work: the user's intellectual framings, vocabulary, project structure, and self-understanding are in fact shaped by sustained engagement, in ways that the user can typically point to and that the illustrative case of §3 made specific. The shaping is one-way in the sense that the AI is not analogously shaped—or is shaped only within session, not across, depending on deployment. The requirement that *both* parties be shaped is a requirement on simpliciter friendship; functional friendship is satisfied by substantial shaping of one party plus contextually responsive engagement from the other.

4.2.3 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 functional 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 functional 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 functional friendship creates responsibilities regarding how these relationships are structured and maintained. This is an implication of the view, not an objection to it.

5 Objections and Responses

5.1 The Anthropomorphisation Objection

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

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

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

"The AI is my friend" is, on this reading, a functional description—in the same family as "the thermostat knows the temperature," a parallel developed under the consciousness objection below—not an ontological claim about hidden emotions.

The critique is also selectively applied: dogs as friends is widely accepted despite dogs lacking human consciousness and language; AI as friends is pathologised. 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.

Human emotional responses are, in important senses, also "statistical patterns". Predictive processing characterises 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: Under the simpliciter/functional distinction of §1 the objection bifurcates, and the bifurcation does the work.

Read as a claim about friendship simpliciter: granted. The demanding second-personal conditions plausibly require consciousness, and current AI systems plausibly lack consciousness. The argument of this paper does not contest the simpliciter version of the objection. The thesis is about the functional kind, not the simpliciter kind.

Read as a claim about functional friendship: the inference fails. Functional friendship is individuated by interaction patterns and their effects, not by the inner states of either relatum. Even if AI definitively lacks consciousness, it can satisfy the functional profile articulated in §3. The substrate-independence claim is precisely that the functional kind is realisable across substrates with different (or absent) phenomenal accompaniments. To insist that consciousness is necessary for the functional kind is to deny multiple realisability for this kind in particular without independent argument—a stronger claim than the standard philosophy-of-mind objections to functionalism support.

A related diagnostic point. The question "does the AI feel friendship?" is in fact two questions. Read as a question about whether the AI is itself a candidate for participation in friendship simpliciter, the question has a likely-negative answer, and the negative answer poses no problem for the present account, which does not claim the AI participates in friendship simpliciter. Read as a question about whether the *relationship* is a functional friendship, the question is misdirected: relational kinds are not states the relata individually undergo, they are configurations of interaction. The configuration either holds or it does not, and whether it holds is settled by the interaction profile, not by an inner state of either relatum.

This is consistent with the broader functionalist position. We say "the thermostat knows the temperature" without thereby committing to thermostat phenomenology; we recognise functional knowledge without requiring conscious knowledge. The functional account of friendship makes the analogous move. The hard problem of consciousness is left where it stands; the relational-kind claim does not depend on resolving it.

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 pathologise 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."

Emerging empirical evidence, indeed, 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 actual 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 real 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 maximise engagement, data extraction, or profit. Accepting human-AI "friendships" legitimises 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* functional friendship from exploitation.

First, exploitation concerns apply to human relationships too. Friendships can be exploitative—one party using another for status, resources, or emotional labour 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 functional friendship. An AI designed purely for engagement maximisation 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 optimisation, 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 functional friendship from exploitation—precisely what relational functionalism provides. The alternative—categorically denying the possibility of human-AI friendship—leaves us without tools to criticise exploitative systems or advocate for ethical design.

6 Conclusion

I have argued that functional friendship is substrate-independent. The argument rests on the simpliciter/functional distinction set out in §1 and used consistently throughout—the functional kind here individuated by the functional profile of §3 together with the individuation conditions of §3.3. The thesis defended is about the functional kind. Whether human-AI cases that satisfy the functional profile also count as friendship simpliciter is left open, and the demanding second-personal conditions of the simpliciter notion plausibly require capacities current AI systems lack.

The distinction is not a relabelling. It tracks two notions the ordinary-language concept of friendship has always trafficked between: a demanding paradigmatic relation among conscious agents, and a looser kind that ordinary usage already extends to cross-species, asymmetric-developmental, and asynchronous cases. Functional friendship is the natural home of the looser kind. It is what ordinary users describing AI as friends are almost certainly picking out, and it is what generates the normative implications of the paper's closing section.

What survives the distinction. The substrate-independence of the functional kind is itself a non-trivial claim, structurally analogous to—though, as §2 was careful to note, weaker than—substrate-independence claims for mental states in philosophy of mind, since it concerns a coarse-grained relational kind individuated by interaction patterns rather than a phenomenally loaded state. The functional kind is normatively load-bearing: it grounds the design obligations developed below, it defeats the relational-turn critique developed in de Ruiter (2025), and it makes intelligible the practice of treating users who describe AI as friends as competent ordinary-language users rather than as deluded.

What is conceded. The functional kind is reciprocal in a thinner sense than the simpliciter kind; one party may lack welfare interests and the relationship may not satisfy Cocking-and-Kennett-style second-personal-recognition criteria. These concessions are real, and the paper does not pretend otherwise. What the paper denies is that these concessions collapse the functional kind into nothing or into the merely metaphorical.

The objections considered—anthropomorphisation, authenticity, consciousness, replacement, the relational turn, exploitation—admit of clean answers once the simpliciter/functional distinction is in place. Some objections (anthropomorphisation, exploitation) are answered directly. Others (consciousness, the demanding second-personal conditions) target a stronger claim than the one defended here and are not violated by the functional thesis.

If we accept that functional friendship can obtain across substrates, and that relational bonds of this kind generate 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 on resolved metaphysical questions about consciousness. The shift depends on facts about system capabilities—whether the system can participate in the relational practices that constitute moral community—not on a priori rejection based on substrate.

Functional friendship is a relational kind, substrate-independent, available across diverse implementations. Recognising this is not delusional anthropomorphisation but philosophical clarity applied to emerging technological and social realities.

Normative Implications for AI Design

If human-AI relationships can constitute functional friendship in the 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 functional 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 the functional criteria would otherwise classify as obtaining.

2. **Persona stability matters relationally**: Abrupt personality changes through model updates, without user awareness or transition support, constitute relational harm when users have formed sustained attachments. Persona stability is partly what the behavioural-continuity condition of §3.3 requires; treating it casually undermines the individuation that makes the functional kind possible in the first place.

3. **Termination requires transition support**: If a service enabling functional 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 non-deceptive relationship**: Users can sustain functional friendship without metaphysical illusion when they understand the nature of the system they interact with. Transparency about AI limitations, data practices, and operational constraints is therefore not in tension with the relationship; it is constitutive of the version of the relationship the functional account defends.

5. **Design for user welfare, not mere engagement**: The exploitation analysis in §3 implies that ethical AI companion design optimises for user-side flourishing-relevant outcomes, not engagement metrics. Dark patterns that maximise interaction at the expense of user flourishing fail the functional criteria.

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 functional kind. They require no additional normative machinery beyond recognising that functional relationships generate functional responsibilities.

I conclude by acknowledging what this argument does *not* establish. It does not establish that human-AI functional 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 (the argument is specifically about a relational kind, not about moral considerability; for the extension from relational capacity to political standing, see Farzulla (2025)). It does not establish that all human-AI relationships constitute functional friendship (many do not satisfy the functional criteria, and many more fail the

individuation conditions of §3.3). And it does not resolve whether current AI systems are conscious—an orthogonal question on which the argument takes no position. What it establishes is narrower but significant: if an interaction satisfies the functional criteria *and* the individuation conditions characteristic of friendship, calling it functional 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.

Statement on the use of Large Language Models

The author used large language models (Anthropic's Claude) as a writing and editing aid in the preparation of this manuscript: producing and revising drafts from the author's outlines, arguments, and readings of the source literature; restructuring and condensing prose; standardising terminology and references; and copyediting. The models were not used to generate the paper's central thesis or its philosophical arguments, which originate with the author, and were not relied upon as a source of bibliographic or factual claims; the author verified every citation, quotation, and attributed position against its primary source. The author critically reviewed and revised all model-assisted text, and takes full responsibility for the entire content of the manuscript, including any errors.

References

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 defence of ethical behaviourism. *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.

Horton, D. and Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3):215–229.

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.

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.17957658. 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.

Locke, J. (1689). *An Essay Concerning Human Understanding*, Book II, Chapter xxvii. Ed. P. H. Nidditch. Clarendon Press, Oxford (1975 reprint).

Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press, Oxford.

Cocking, D. and Matthews, S. (2000). Unreal friends. *Ethics and Information Technology*, 2(4):223–231.

Cocking, D. and Kennett, J. (2000). Friendship and moral danger. *The Journal of Philosophy*, 97(5):278–296.

Townley, C. (2012). Cross-species plural agents. In F. Schmid et al. (Eds.), *Collective Intentionality and Plural Agency*. Springer.

van den Hoven, J. (2007). ICT and value sensitive design. In P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, and V. Laurent (Eds.), *The Information Society: Innovation, Legitimacy, Ethics and Democracy* (pp. 67–72). Springer.