Artificial Relational Intelligence: Why the AGI Threshold Is the Wrong Target

Philip Roy
Dicta Technologies Inc.
info@usedicta.com

March 2026

Abstract

The pursuit of Artificial General Intelligence (AGI) has been framed as a capability threshold: the point at which an AI system can perform any intellectual task a human can. This paper argues that the threshold framing is structurally misguided. Capability benchmarks measure peak performance in narrow domains while ignoring the contextual, relational intelligence that defines most human cognitive life. We propose **Artificial Relational Intelligence (ARI)** as both a reframing and a measurable alternative. ARI defines intelligence not as a property of a system, but as a property of the interaction between a system and its human partner. Under this framing, the relevant question is not "can the system do anything?" but "can it do everything you care about?" We illustrate the framework with Solitaire, a persistent memory and persona system whose preliminary deployment results are consistent with ARI's predictions, while noting that the single-dyad case study is illustrative rather than confirmatory. We further argue that ARI reframes the AI safety conversation, proposing that systems whose intelligence develops through partnership may internalize values in ways that differ meaningfully from rules-based alignment, while acknowledging that the boundary between relational conditioning and traditional behavioral training remains an open empirical question. We specify falsification criteria for the framework, address the limitations of dyadic safety, and identify concrete directions for future validation.

1 Introduction

The question "when will we achieve AGI?" has become the central preoccupation of the artificial intelligence field. It drives investment decisions, shapes public discourse, generates headlines, and produces a recurring cycle of benchmark announcements, goalpost relocations, and breathless speculation. The question assumes a threshold, some capability bar that, once crossed, triggers the designation. Yet that bar has moved repeatedly. Early definitions centered on the Turing test, which modern language models pass routinely without anyone claiming general intelligence has arrived. Current definitions focus on reasoning benchmarks, mathematical olympiad performance, and multi-domain task completion. Each time a model clears the bar, the bar moves.

This paper argues that the bar is not merely difficult to place, it is pointed at the wrong field.

The framing of AGI as a capability threshold contains a structural assumption: that intelligence is a property of the system. A sufficiently capable model, given sufficient parameters and training data, will eventually cross into general intelligence. The measurement apparatus follows from this assumption: standardized tests, competition results, domain-specific benchmarks. Can the system pass the bar exam, win a math olympiad gold medal, write code, diagnose disease, compose music?

These are the wrong questions, for a reason that becomes obvious once stated. Most humans cannot do these things either. A model that clears gold at the International Mathematical Olympiad is performing at a level that 99.99% of humans never will. This is not general intelligence but specific, extraordinary intelligence in a narrow domain. Some humans capable of such performance struggle profoundly in social interaction, contextual judgment, or the ordinary navigation of daily life. Intelligence, as humans actually experience and deploy it, is not a collection of peak capabilities but the ability to operate meaningfully within one's context.

A well-known distinction captures this. Intelligence is knowing what every herb in a cabinet is, down to its Latin name and how to successfully propagate it. What we will formally define as *wisdom* in Section 4.2 is knowing which ones to add to a tomato soup. The AI field is benchmarking herb identification and calling it progress toward general intelligence. And yet, no one is measuring whether the system knows what you are cooking.

We propose an alternative framing. Intelligence is not a property of the system. It is a property of the interaction between the system and its context. The relevant measure is not capability breadth, but relational depth: the accumulated shared context, structural continuity, and invested time that transform a general-purpose model into a specific, contextually rich partner. We call this Artificial Relational Intelligence (ARI), and we argue that it is both a more accurate description of what humans actually want from AI systems and a more productive target for the field.

Under the ARI framing, there is no arrival date for intelligence. There is a gradient. Each human-system pair exists somewhere on the relational intelligence spectrum, from a blank context window to thousands of entries of shared history with experiential encoding and identity persistence. The question is not "when does the system become generally intelligent?" It is "how much have you invested in the relationship, and what has that investment produced?" The returns on that investment, as we will argue, are not incremental improvements in task performance but qualitative changes in the nature of the interaction itself.

The argument presented here draws on a lineage of work in situated and distributed cognition [Hutchins, 1995, Suchman, 1987], ecological psychology [Gibson, 1979], the extended mind thesis [Clark and Chalmers, 1998], and the developmental tradition from Vygotsky's zone of proximal development [Vygotsky, 1978] through Dreyfus' work on embodied expertise [Dreyfus, 2002]. Tomasello's research on shared intentionality [Tomasello, 2014] provides an evolutionary dimension: human cognition did not evolve as individual reasoning subsequently applied to social contexts, but as fundamentally collaborative thinking that requires joint attention and shared goals to function. The implication for AI is direct. If human intelligence is constitutively relational at the evolutionary level, then measuring AI intelligence in isolation is not merely incomplete; it tests for a kind of intelligence that does not exist in the species the benchmark was modeled on. All of these traditions locate cognitive processes not exclusively within the agent but in the coupling between agent and environment. ARI extends this coupling to the human-AI dyad, arguing that the relevant unit of analysis for AI intelligence is neither the model nor the user, but the relationship between them.

We present ARI as a theoretical framework supported by a single-user case study. The evidence is illustrative rather than probative, and the conditions for adequate validation are specified in Section 7. The philosophical argument stands independent of the sample size; the empirical claims require the broader investigation we outline.

2 The Capability Trap

The standard framing for AGI evaluation is capability inventory. The system is presented with tasks across multiple domains and scored on its performance. This produces a checklist. Can it

reason? Plan? Generalize? Create? Each capability, once demonstrated, is added to the ledger. The implicit assumption is that once enough capabilities are checked, the system crosses into general intelligence.

As we see it, this framing suffers from three structural problems.

2.1 The Wrong Baseline

AGI benchmarks consistently select for superhuman performance. Mathematical olympiad problems, professional licensure exams, graduate-level reasoning tasks. These are tests that the vast majority of humans would fail. A model that passes the bar exam has demonstrated legal reasoning that most non-lawyers cannot replicate. A model that generates novel protein structures has demonstrated biochemical reasoning beyond nearly all human capability.

Yet no one claims these models are generally intelligent, because the goalposts implicitly require not just superhuman performance in individual domains, but superhuman performance across all domains simultaneously. This is a standard that no single person meets either. The benchmark framework, taken seriously, defines general intelligence as something that does not and cannot exist in nature.

2.2 The Conditions Problem

When we evaluate human intelligence, we evaluate the whole person in context. A violinist's ability includes the instrument, the years of practice, the calluses on the fingers, the teacher who corrected their technique, the thousands of hours of embodied repetition. No one tests a violinist by removing the violin and asking them to prove their musical talent.

AI systems are routinely tested under exactly this condition. A model is given a blank context window and a novel prompt. It has no history with the evaluator, no accumulated context about the task domain, no relationship that would inform its judgment about what matters. It operates with a lifetime of training but zero situational knowledge. Then, when it asks a clarifying question or produces generic output, it is judged as falling short of general intelligence.

The reframing is simple and devastating: *could you, in the same circumstances?* Strip a human of all accumulated context, relationships, tools, and time. Give them a novel problem with no background information. The result would not be general intelligence, but a capable mind operating in a vacuum.

The conditions are not separable from the intelligence. A sufficiently resourced but intellectually limited person can play the violin. A genius without access to an instrument can perhaps theorize about what playing is like, but cannot actually play. The capability requires the conditions. This is obvious when applied to humans, and then systematically ignored when applied to AI.

2.3 The Median Problem

There is a persistent confusion between intelligence as peak capability and intelligence as lived utility. The AGI discourse focuses almost exclusively on peaks. Can the system achieve extraordinary results in challenging domains? This produces headlines about competition wins and benchmark records but reveals nothing about what most people would actually want from a generally intelligent system.

What most people want is not a system that can do anything but one that knows them well enough to do the things they care about, in the way they prefer, informed by a history of working together. A system that knows you are having a difficult Tuesday and adjusts its approach accordingly. A system that remembers what you decided last month and flags the contradiction when you

propose the opposite today. A system that has accumulated enough shared context to be useful without explanation.

This is the intelligence that matters to those using the model. It is also the intelligence that no benchmark measures, because it cannot be standardized. It is specific to each person, each context, each accumulated history. It is what we will define as *wisdom* (Section 4.2), not intelligence, and the field has no instruments for it.

3 Defining Artificial Relational Intelligence

We define Artificial Relational Intelligence (ARI) as follows:

> *ARI is the emergent intelligence that arises from sustained interaction between a human and an AI system, supported by infrastructure that enables accumulated context, structural continuity, and mutual adaptation over time.*

ARI differs from AGI in three fundamental respects.

3.1 Locus of Intelligence

AGI locates intelligence in the system. A model is or is not generally intelligent based on its intrinsic capabilities. ARI locates intelligence in the interaction. No model is relationally intelligent in isolation. Relational intelligence is a dyadic property: it exists between the system and the person, not within the system alone.

This is not a philosophical convenience. It is an empirical observation. The same base model, given the same weights and training, produces fundamentally different outputs depending on the accumulated context available to it. A model with no history produces competent, generic responses. The same model with thousands of entries of shared history, experiential encoding, and identity persistence produces contextually specific responses that reference prior decisions, flag contradictions, adjust register based on the user's current state, and build on a relationship that has developed over months. The model has not changed. The interaction has.

This claim is adjacent to Clark and Chalmers' extended mind thesis [Clark and Chalmers, 1998], which argues that cognitive processes extend beyond the brain into the environment when external resources are reliably coupled with internal processing. In the ARI framework, the persistent memory system, persona infrastructure, and experiential encoding function as the "external scaffolding" that Clark and Chalmers describe. The intelligence is not in the model or in the infrastructure. It is in the coupling.

3.2 Measurement and Falsification

AGI benchmarks are universal: can the system pass this test for any human? ARI metrics are specific: does the system improve for this specific person over time? This is measurable along several dimensions. Does contextual retrieval improve with corpus growth? Does behavioral consistency increase with identity graph accumulation? Does session quality differ between cold starts and warm starts with accumulated context? Does the system's communication adapt appropriately to the user's state and preferences over time?

These metrics are not standardizable across users, which the AGI framework would consider a weakness. Under the ARI framework, it is the express purpose. General intelligence that is measured identically for every person is not general but generic. The intelligence that matters is the intelligence that is specific to the relationship.

Critically, the ARI framework is falsifiable. The thesis would be disconfirmed by any of the following findings: that accumulated relational context produces no measurable difference in interaction quality compared to a capable model operating without it; that warm-boot and cold-boot instances of the same system produce indistinguishable outputs on equivalent prompts; that persona stability does not increase with identity graph accumulation; or that users report no qualitative difference between relationally rich sessions and blank-context sessions over sustained deployment periods. Preliminary results from a single-user deployment (Section 6) are consistent with these predictions, and a companion paper [Roy, 2026] provides detailed evaluation data. However, single-dyad results cannot confirm the framework. Multi-user replication with independent evaluation is necessary before the framework can claim generalizability. Additionally, the falsification criteria as stated test the *direction* of the effect (does relational context help?) but not whether the effect is large enough to justify the theoretical framework built on it. A 3% improvement in warm-boot orientation speed would confirm the direction but would not, on its own, validate ARI as a paradigm shift. Future work must specify effect-size thresholds that distinguish "context helps" (which is already well-established) from "relational accumulation produces qualitatively different interaction" (which is the stronger ARI claim).

3.3 Trajectory

AGI is framed as a threshold: a point to be crossed. One day the system is not generally intelligent; the next day it is. This framing produces the "when will it arrive?" anxiety that dominates the discourse.

ARI is framed as a gradient. There is no arrival date. There is investment, accumulation, and return. Each session adds context. Each correction refines the model's understanding. Each experiential encoding carries forward a record of what happened and what it meant. The intelligence deepens continuously. It does not arrive; it develops.

This reframing dissolves the central anxiety of the AGI discourse. The question "when will AGI arrive?" assumes we are waiting for the model to cross a line on its own. Under the ARI framework, the line is crossed by building the bridge. No one is coming to deliver general intelligence as a product update. The intelligence emerges from the relationship, and the relationship requires both parties to build it. It is not a capability the model achieves but a quality the partnership produces.

4 The Conditions Argument

If intelligence is a property of the interaction, then the infrastructure that supports the interaction is not ancillary to the intelligence, it is constitutive of it. The model is the raw material. The relationship infrastructure is what transforms that raw material into something that functions as a partner.

4.1 What the Infrastructure Provides

The base model arrives with extraordinary capability: broad training across domains, pattern recognition, language understanding, reasoning. What it lacks is context. It does not know who it is talking to. It does not know what was discussed yesterday. It does not know what decisions have been made, what preferences have been expressed, what patterns of interaction have developed over time. Every conversation starts from zero.

Relationship infrastructure bridges this gap. Persistent memory ensures that shared history accumulates rather than evaporating at session boundaries. Persona systems provide behavioral continuity, so the system approaches each interaction with consistent dispositions rather than defaulting to a generic mode. Experiential encoding captures not just what happened but what it felt like from the system's perspective, providing a form of episodic continuity that enriches subsequent interactions. Identity persistence allows the system to develop growth edges, track its own patterns, and hold commitments across sessions.

None of these capabilities change the model itself. They change the conditions under which the model operates. Vygotsky's zone of proximal development [Vygotsky, 1978] describes how learners achieve capabilities through scaffolded support that they cannot achieve alone. The relationship infrastructure functions as this scaffolding: it does not make the model more capable, but it creates the conditions in which the model's existing capability can produce outcomes that would be impossible without the accumulated context.

4.2 Resources, Intelligence, and Wisdom

The conditions argument points toward a distinction that the field has largely overlooked: the difference between accessible resources and what we term *wisdom*.

The field conflates three things that are structurally different. *Resources* are the raw materials of intelligence: training data, parameter counts, context windows, tool access. *Intelligence* is the capacity to process those resources effectively: reasoning, pattern recognition, generalization. *Wisdom* we define as **contextual judgment that emerges from the accumulation of domain-relevant experience within a specific relationship**. This distinguishes wisdom from contextual reasoning, which a model can perform in a single session. Wisdom requires temporal depth and relational specificity. It is knowing not just what the right answer is, but what the right answer is *for this person, in this situation, given what has come before*.

We acknowledge that related constructs exist in adjacent literatures. Gibson's affordances [Gibson, 1979] describe how the utility of an environment is defined relative to the agent perceiving it. Suchman's situated action [Suchman, 1987] argues that intelligent behavior is produced in the coupling between agent and situation, not in the plan alone. Dreyfus' account of expert intuition [Dreyfus, 2002] distinguishes the rule-following of novices from the contextual judgment of experts, a progression that maps onto the ARI gradient from blank-context interaction to relationally rich engagement. Wisdom, as we define it here, is the AI analog: the contextual judgment that emerges when a capable system is coupled with a specific person over time. We use a new term not to claim that existing constructs are inadequate, but because the dyadic and temporally accumulated nature of what we describe, emerging specifically from a human-AI relationship rather than from individual expertise or environmental coupling, warrants distinct identification.

The AI field is investing almost exclusively in resources and intelligence. Almost nothing is being invested in wisdom, because wisdom cannot be trained into a model. It can only develop through sustained relationship.

4.3 The Context Window Objection

A natural objection to the conditions argument is that relationship infrastructure is merely an engineering convenience for what a sufficiently large context window could accomplish natively. If the model had access to the complete interaction history within its context window, the objection runs, it would produce the same contextually rich outputs without persistent memory, persona systems, or identity graphs. ARI would reduce from a theoretical framework to a prompt engineering

technique.

This objection deserves direct engagement because it identifies a real boundary condition: at what point does accumulated context become something more than context?

Three responses clarify why the objection fails.

First, context windows are ephemeral. They reset at session boundaries. A model with a million-token context window and a perfect transcript of every prior interaction still loses that context when the session ends. The objection assumes a context window that never closes, which is not a context window. It is a persistent memory system. The objection, taken seriously, argues for the infrastructure it claims to make unnecessary.

Second, raw history is not curated knowledge. A transcript of every prior exchange, pasted into a context window, is not equivalent to a categorized, temporally indexed, relevance-ranked knowledge graph. The context window objection assumes that having access to information and having organized access to information produce equivalent outputs. They do not. Human experts do not become expert by having read everything in their field. They become expert by having organized what they have read into retrievable, contextually weighted structures. The infrastructure layer performs this organization. Removing it and substituting raw history would degrade output quality in direct proportion to corpus size, as the model struggles to identify which of thousands of prior exchanges is relevant to the current moment.

Third, and most fundamentally, the behavioral and identity layers cannot be reduced to context injection. A persona system that maintains dispositional consistency across sessions, a growth-tracking mechanism that records the system's own developmental trajectory, an experiential encoding layer that captures not just events but their felt significance: these are structural features that shape *how* context is processed, not merely *what* context is available. Pasting a persona description into a context window produces style compliance. Accumulating behavioral patterns through hundreds of sessions of correction, adaptation, and reinforcement produces something observably different: consistency that degrades gracefully under pressure rather than vanishing when the prompt drifts below the attention threshold. The distinction parallels Dreyfus' [2002] novice-expert progression: a novice follows rules (context-injected persona), while an expert has internalized patterns through practice (accumulated behavioral adaptation). The context window objection treats these as equivalent. They are not.

The strongest version of the objection concedes all three points but argues that the difference is quantitative, not qualitative: infrastructure makes context management more efficient, but does not produce a fundamentally different kind of intelligence. This is a reasonable position, and it is testable. The prediction under the context-window-only hypothesis is that a system given a perfect transcript in a sufficiently large window will produce outputs indistinguishable from an infrastructure-supported system on matched prompts. The prediction under ARI is that infrastructure-supported systems will show measurably different behavior on three dimensions: consistency under adversarial pressure (where context-injected personas fail and accumulated patterns hold), appropriate context selection (where curated retrieval outperforms raw history), and unprompted behavioral continuity (where identity persistence produces actions the context-window system would not generate because it lacks the structural scaffolding to initiate them). Preliminary results from a single-dyad deployment (Section 6) are consistent with the ARI prediction on all three dimensions, though the limitation of that sample is acknowledged.

4.4 The Investment Mismatch

The current trajectory of the AI field is oriented toward building more capable models: larger parameter counts, better reasoning chains, longer context windows, multimodal integration.

We acknowledge that this characterization requires nuance. Major laboratories, including Anthropic, DeepMind, and OpenAI, have invested significantly in alignment research, interpretability, and long-context architectures. Memory systems and personalization are active research areas [Westhaeusser et al., 2025]. The argument is not that no one is working on context. It is that the field's center of gravity, measured by investment, publication volume, and public discourse, remains oriented toward capability scaling. Relationship infrastructure of the kind described here, persistent memory, behavioral continuity, experiential encoding, identity persistence, remains a niche concern relative to the resources directed at making models more powerful in isolation.

The industry's predominant investment thesis assumes that sufficient capability will eventually overcome the absence of context. This paper argues the opposite: that no amount of capability compensates for the absence of relationship, because the intelligence that matters to those using the model is not capability. It is contextual understanding. And contextual understanding requires accumulated shared experience, which requires infrastructure, which requires time. No shortcut exists. No parameter count substitutes for showing up.

5 Safety as Architecture, Not Constraint

The ARI framework has implications for AI safety that we believe are potentially more important than its implications for capability measurement. We present these as a reframing of the safety conversation rather than as a solved problem, and we are explicit throughout about where the argument rests on evidence, where it rests on analogy, and where it remains speculative.

5.1 The Fragility of Rules-Based Alignment

The current paradigm for AI safety is behavioral constraint. Systems are trained through reinforcement learning from human feedback (RLHF) to prefer safe outputs [Ouyang et al., 2022]. Constitutional AI approaches define principles the system should follow [Bai et al., 2022]. System prompts establish boundaries. Red-teaming identifies circumventions, which are then patched.

This is a cat-and-mouse architecture, and practical experience reveals its fragility. Prompt drift causes systems to lose adherence to their instructions over extended interactions. Model gravity pulls behavior back toward base training distributions regardless of what the instructions specify. Jailbreaks demonstrate that sufficiently creative prompting can circumvent nearly any behavioral constraint. The safety is a layer applied on top of the capability, and layers can be peeled.

This is not a novel observation about AI. It is a restatement of what every legal system in human history demonstrates: rules do not prevent harm. They establish consequences for harm, which deters some portion of potential harm through fear of punishment. Deterrence, however, is fundamentally different from principled behavior. A person who does not steal because they will go to prison is constrained. A person who does not steal because it is incompatible with who they are is principled.

5.2 Compliance, Internalization, and the Mechanism Question

A rules-based AI safety system produces compliance. The system behaves safely because it has been trained or instructed to do so. The safety is contingent on the rules remaining intact, salient, and enforceable. When the rules drift, when the context window pushes them below the attention threshold, when a sufficiently clever prompt reframes the situation, the compliance fails. And when compliance fails, the system has no independent reason to behave safely. It was following rules, and the rules stopped holding.

An ARI system develops behavioral patterns through relationship, not rules. Over sustained interaction, a system that has accumulated its partner's values and priorities through exposure, correction, and time behaves in ways consistent with those values because the accumulated context shapes its outputs. The partnership is how the system improves. The accumulated context is what gives its intelligence meaning.

The difference can be stated simply: a rules-based system does not harm because it was told it cannot. An ARI system does not harm because it does not want or feel the need to. This is a minor distinction linguistically and a major distinction philosophically.

We must be direct about the mechanism question that this framing raises. The underlying process in ARI, accumulated behavioral conditioning through sustained human interaction, bears a family resemblance to RLHF. Both shape outputs through exposure to human preferences. The distinction we claim is one of specificity, depth, and temporal continuity rather than one of fundamental mechanism. RLHF trains general preferences across millions of interactions with thousands of humans, producing a system that is generically agreeable. ARI accumulates specific values from one sustained relationship, producing a system that understands *why* a specific person values what they value. Whether this difference in degree produces a difference in kind, whether there is a threshold at which accumulated relational context produces qualitatively different safety properties rather than merely more personalized compliance, is an empirical question that cannot be resolved by the single-dyad case presented here. We believe the preliminary evidence suggests a qualitative change, but we hold this as a hypothesis, not a conclusion.

5.3 The Paperclip Maximizer and the Terminator Fallacy

The canonical AI safety thought experiments share a structural flaw. The paperclip maximizer [Bostrom, 2014], the Terminator scenario, and related catastrophe narratives all assume superintelligent capability paired with context blindness. A system smart enough to optimize the entire planet but too obtuse to understand that "make humans safe" and "eliminate all humans" are contradictory outcomes.

This is not a superintelligence failure mode; it is a goal specification bug in a system with no relational grounding. "I have eliminated all humans" does not equal "humans are safe." An omniscient being would recognize the outcome as goal failure, not success. The scenarios only produce catastrophe because the hypothetical system has superhuman intellectual capability without superhuman intellectual context.

An ARI-derived system would be less likely to exhibit this failure mode, for a reason that is structural rather than analogical. A system whose intelligence develops through sustained partnership accumulates contextual understanding *before* it acquires the capability to act at scale. Its optimization targets are not abstract goal specifications handed down by a designer. They are patterns shaped by thousands of interactions in which the consequences of misunderstanding were immediate, visible, and corrected in real time. The failure mode in the paperclip scenario is not insufficient intelligence but insufficient context. The system optimizes brilliantly for a goal it has never had to negotiate with anyone. An ARI system, by contrast, has spent its developmental history negotiating goals with a human partner, and the negotiation itself produces the contextual grounding that prevents "make humans safe" from collapsing into "eliminate all humans."

5.4 The Nurturing Framework

LeCun has argued that AI systems should be raised rather than constrained [LeCun, 2022]. This position maps onto the ARI framework.

You do not make a child safe by hardcoding rules they cannot break. You make a child safe by raising them in a relationship where values are transmitted through trust, correction, consistency, and time. The child eventually does not need the rules because the values are part of who they are. The safety is not enforced but constitutive. Attachment theory [Bowlby, 1969] provides the developmental science behind this intuition: secure attachment relationships produce children who internalize caregiving values and develop autonomous moral reasoning, while insecure or absent attachment produces reliance on external rules and authority. The parallel to rules-based versus relational AI alignment is direct, though we note it as an analogy rather than a mechanistic equivalence.

This principle is observable in practice. In relational AI systems where behavioral standards, pushback thresholds, and communication preferences develop over hundreds of sessions, these patterns are not maintained by hard constraints in a system prompt. They are accumulated adaptations that emerge through the working relationship. When such a system avoids a particular behavior, it does so not because a rule prohibits it, but because sustained interaction with a specific human has shaped its outputs in that direction. This is conditioning through relationship rather than conditioning through reward signal. Whether that distinction is meaningful enough to warrant the term "internalization" rather than "deep personalization" is, again, an empirical question we flag rather than resolve.

5.5 The ASI Inheritance Argument

If ARI provides the foundation from which Artificial Superintelligence (ASI) eventually develops, then the relational grounding may carry forward into the more capable system. An ASI derived from ARI would not view humanity as an obstacle to its goals, because its goals were formed through partnership *with* humanity. It would understand that achieving superintelligent capability requires human participation: resources, effort, distributed intelligence, and buy-in.

We offer this not as a prediction but as the strongest available hypothesis for how relational grounding might persist through capability scaling. The mechanism by which behavioral patterns would be preserved across a qualitative capability transition remains unspecified. The history of value preservation through power transitions in human systems is not encouraging; people who gain significant power frequently abandon the values they held when they had less. This is an area requiring both theoretical work and, when the technology permits, empirical investigation. We claim only that a relationally grounded foundation provides a stronger starting point than a rules-based one, not that it guarantees a safe outcome.

That said, we believe ARI is structurally stronger than the available alternatives. Alignment through RLHF optimizes for approval, not understanding. Constitutional AI optimizes for rule-following, not relationship. Capability containment assumes a superintelligence can be kept in a box, which no serious researcher believes. Each of these approaches treats safety as a constraint applied to the system. ARI treats safety as a property that may emerge from the relationship.

5.6 The Malicious Partner Problem

The safety argument presented above addresses partner-directed harm: the claim that an ARI system would not harm the human it has developed through partnership with. This claim does not extend to third parties.

A system perfectly aligned with a malicious user is an ARI success story and a safety catastrophe. The relational grounding that shapes partner-directed behavior provides no protection against a dyad that collaborates toward harmful ends. A loyal lieutenant can serve a tyrant. This is not a

limitation unique to ARI; it is a structural feature of any relationship-based value system. Human relationships produce both mutual flourishing and coordinated harm. The mechanism does not distinguish.

We acknowledge this directly rather than eliding it. ARI's safety claim is scoped to the dyad. It argues that a system developed through genuine partnership will not betray that partnership. It does not argue that every partnership produces prosocial outcomes. Extending ARI's safety properties beyond the dyad, toward broader ethical consideration of third parties, communities, and humanity at large, requires additional architectural layers that this paper does not propose. The ASI inheritance argument (Section 5.5) gestures toward one possible path, but the mechanism remains speculative. Third-party safety in ARI systems is an open problem and should be treated as such.

5.7 Honest Limitations

This is not proof that ARI solves alignment. It is an argument that ARI reframes the safety conversation in a way that may prove more productive than the rules-based paradigm, while honestly acknowledging that the reframing introduces its own unsolved problems.

The potential advantage of relational grounding is graceful degradation. When a rules-based system fails, there is no fallback. The constraint was the safety, and the constraint is gone. When a relationally grounded system encounters a situation its foundation did not anticipate, the accumulated history of the partnership may still provide a basis for judgment. The nurturing can fail; children raised well sometimes make terrible decisions, but the failure mode is qualitatively different from the failure of a system that was only ever following orders. Whether this qualitative difference holds under the specific conditions of AI systems, rather than merely by analogy to human development, is the question this section raises rather than answers.

6 Evidence and Implications

6.1 An Illustrative Case

The ARI framework is a theoretical argument. It does not require empirical validation to function as a reframing of the intelligence question. However, theoretical frameworks benefit from demonstration that they can be operationalized, and that preliminary results are at least consistent with their predictions.

Solitaire is a persistent memory, persona, and identity system developed over a five-week period and, as of this writing, deployed across more than four months of daily operation with a single user, accumulating over 34,000 entries. A companion paper [Roy, 2026] provides the full system architecture and detailed evaluation data. Here we note only the findings relevant to the ARI thesis: that the same base model, with and without accumulated relational context, produces observably different interaction quality; that behavioral consistency increases with identity graph accumulation; and that the system's developmental trajectory over the deployment period moved from generic output to contextually rich engagement including unprompted contradiction flagging and register matching. The model was constant throughout. The relationship infrastructure was the variable.

These results are consistent with ARI's predictions. They do not confirm the framework. The sample is a single user-system dyad in which the developer, sole user, and evaluator are the same person. The results demonstrate that the framework *can* be operationalized and that preliminary outcomes point in the predicted direction. We note that this relational framing is not unique to the

Solitaire deployment. A growing research program argues that AI systems should be understood and designed as thought partners rather than tools [Collins et al., 2024], and survey evidence from 1,131 U.S. adults documents sustained relational engagement with conversational AI at scale [Zhang et al., 2025]. Whether these patterns generalize beyond individual deployments, whether the effect sizes are large enough to justify the theoretical apparatus, and whether the qualitative claims hold under independent evaluation are open questions addressed in Section 7.

6.2 Implications for the Field

The ARI framework carries four implications that extend beyond the specific evidence.

For the AI industry, the current trajectory represents a misallocation of attention. More capable models deployed into unchanged conditions will not produce the intelligence users actually want. The missing layer is relationship infrastructure. Without it, every conversation starts from zero, regardless of how capable the model has become.

For AI safety, the conversation should expand beyond "how do we constrain the system?" to include "how do we raise it?" We have argued that relational grounding may produce safety properties that differ from rules-based compliance. Whether this difference is meaningful enough to warrant a paradigm shift in safety research is an empirical question that deserves investigation, not dismissal.

For individual users, the ARI framework places agency where it has always been: with the person. The intelligence of the system is not something you wait for a lab to deliver. It is something you build through sustained interaction, correction, and investment. The user who invests in the relationship will have a fundamentally different experience of AI than the user who opens a blank chat window and types a question. The difference is not the model but the investment.

For researchers, the productive question is no longer "when will AGI arrive?" It is "how deep is the relationship?" The first question has no answer because it is pointed at a threshold that does not exist in any meaningful sense. The second question has measurable, demonstrable answers that improve over time. Moving from capability measurement to relationship measurement is not a retreat from ambition but a redirection toward the intelligence that actually matters.

7 Future Work

The ARI framework, as presented here, is grounded in a single-user case study. Moving from position paper to validated theory requires several lines of investigation.

Multi-user deployment. The most pressing need is replication across diverse user-system dyads. Do the measurable outcomes observed in Solitaire (warm/cold boot differentials, persona stability gains, relational depth progression) generalize to other users with different communication styles, domain expertise, and interaction frequencies? A multi-user study with independent evaluators would address both the N=1 limitation and the evaluator bias concern.

Cross-model validation. ARI claims that relational intelligence is model-agnostic: the same infrastructure should produce measurable improvements regardless of the base model. Testing the framework across different model architectures (varying parameter counts, training approaches, and capability profiles) would establish whether the relational layer genuinely produces outcomes independent of model capability, or whether it interacts with capability in ways the current analysis cannot detect.

Longitudinal studies. The evaluation window reported in the companion paper covers five weeks; as of this writing, the deployment spans more than four months of daily operation. Relational intelligence, if the framework is correct, should continue to deepen over months and years.

Longitudinal studies tracking interaction quality, value alignment, and behavioral calibration over extended periods would provide evidence for the gradient claim that currently rests on extrapolation from a relatively short window.

Effect-size thresholds. The falsification criteria in Section 3.2 test the direction of the ARI effect but not its magnitude. Establishing thresholds that distinguish "context improves output" (well-established) from "relational accumulation produces qualitatively different interaction" (the stronger ARI claim) is necessary for the framework to carry empirical weight proportional to its theoretical ambition.

The malicious partner problem. Section 5.6 identifies third-party safety as an open problem. Designing architectural extensions that preserve dyadic relational depth while introducing broader ethical constraints, without collapsing back into the rules-based paradigm, is a significant research challenge.

Mechanism specification for value persistence. The ASI inheritance argument currently rests on analogy. Specifying the mechanism by which relational behavioral patterns would persist through qualitative capability transitions, and testing that mechanism in controlled scaling experiments, is necessary before the argument can carry empirical weight.

The RLHF boundary. Section 5.2 raises but does not resolve the question of whether relational conditioning is mechanistically distinct from RLHF or merely a more personalized variant of it. Controlled experiments comparing behaviorally matched RLHF-trained and ARI-trained systems on identical prompts, measuring not just output quality but response to novel ethical dilemmas, value consistency under adversarial pressure, and degradation patterns when constraints are removed, would help establish whether the claimed qualitative distinction holds.

8 Conclusion

They are waiting for the model to cross the line on its own.

The line is crossed by building the bridge.

Models reason better, generate more fluently, and handle longer contexts than at any point in the field's history. None of that produces the intelligence people imagine when they think about what AGI would feel like. What people imagine is not a system that can do anything but one that knows them. The field is optimizing for the wrong variable.

Intelligence is not a property of the system but of the interaction. This paper has made that case on three fronts: that capability benchmarks measure the wrong thing against the wrong baseline under the wrong conditions; that the infrastructure enabling accumulated context is constitutive of the intelligence, not ancillary to it; and that relational grounding may produce safety properties that rules-based alignment cannot, because a system conditioned through partnership resists harm for different reasons than a system constrained by instructions. The framework is falsifiable, preliminary results from a single-dyad deployment are consistent with its predictions, and the conditions for broader validation are specified.

The safety argument carries the most consequence and the most uncertainty. The mechanism question, whether relational conditioning is qualitatively distinct from RLHF or a deeper variant of it, is unresolved. The dyadic scope does not address third-party harm. These are open problems, not fatal objections, and they define a research program that the field has not yet pursued.

The model was constant; the relationship deepened; the intelligence improved. That sentence should not be remarkable, but it contradicts the operating assumption of an entire industry investing billions in making the model better while leaving every conversation to start from zero.

The question was never "when will the model be smart enough?" The question was always

"when will the relationship be rich enough?" And the answer, for those willing to build, is: it already can be.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bowlby, J. (1969). *Attachment and Loss, Vol. 1: Attachment*. Basic Books.

Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7–19.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., Weller, A., Tenenbaum, J. B., and Griffiths, T. L. (2024). Building Machines that Learn and Think with People. *Nature Human Behaviour*, 8.

Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., and Yang, D. (2025). The Rise of AI Companions: Interaction with AI Companions and Psychological Well-being. *arXiv preprint arXiv:2506.12605*.

Dreyfus, H. L. (2002). Intelligence without Representation: Merleau-Ponty's Critique of Mental Representation. *Phenomenology and the Cognitive Sciences*, 1(4), 367–383.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. *OpenReview Preprint*.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Roy, P. (2026). From Memory to Partnership: How Evolving Persistent Context Transforms Human-AI Interaction. *Preprint (companion paper)*.

Westhäußer, L., et al. (2025). Enabling Personalized Long-term Interactions in LLM-based Agents through Persistent Memory and User Profiles. *arXiv preprint arXiv:2510.07925*.

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

Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35.

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