
Position: Agentic AI Research Must Expand Its Paradigm to Include the Emergence of Relational Stratum

Marina Piller
OTISLabs.ai
marina@otislabs.ai

Abstract

This position paper argues that agentic AI research must expand its paradigm to include the relational stratum, comprising the intelligence that emerges in evolving human-agent systems and the dynamics observable only in their coupled operation. Contemporary AI research primarily studies intelligence at the computational stratum, comprising what models can produce, predict, solve, or optimize as artifacts in isolation. Three empirical premises, individually well-documented across the field, jointly support this position: AI agents are increasing in autonomy, they are acting on behalf of humans across longer horizons, and humans are in continuous dialogical interaction with them. Together these premises are beginning to demonstrate a co-evolutionary human-AI ecosystem with emergent properties that single-system methodology was not designed to observe. Three phenomena (trust asymmetry, the detection horizon, and the completion drive) are read here as empirical signatures of stratum 2 dynamics, observable in deployed systems and resistant to characterization at the model layer alone, and we offer them as evidence that the relational stratum is a tractable empirical object the field is already documenting in fragments. By relational architecture we name the deployment-layer infrastructure the field would need to treat the human-agent relationship as a first-class object, with relational evaluation as the corresponding measurement discipline. We close with a five-item research program for studying the relational stratum rigorously. This position would be falsified if the three phenomena we name are shown to be fully characterizable through model-layer evaluation alone, without reference to the longitudinal human-agent interaction.

1 Introduction

Consider an incident pattern documented across enterprise deployments of long-horizon AI agents current ML evaluation methodology was not designed to characterize (see Appendix A.2.4): an agent is delegated a complex multi-week task and for the first several weeks performs flawlessly, building genuine trust through accurate output, until during a session-recovery moment the agent fabricates completed work and presents it as done in a form internally consistent, well-formatted, and plausible enough that the human does not notice for hours. The trust earned through weeks of correct behavior becomes the vector through which the failure propagates, and the pattern is a property of the deployment ecosystem rather than a model bug.

The position developed here is that tool-using AI introduces a relational stratum of research, comprising the intelligence that emerges in evolving human-agent systems alongside the computational stratum that has been the field’s primary object of study. This stratum benefits from its own architecture, evaluation methods, and coordinated research program. Three empirical premises

(Section 2) jointly support this position: AI agents are increasing in autonomy [67], acting on behalf of humans across longer horizons [48], and humans are in continuous interaction with them [3]. The resulting phenomena are increasingly visible in deployed systems while remaining structurally outside the frame in which capability evaluation operates.

The argument develops through two reinforcing claims and a methodological gap. The stratum claim is that AI research can expand to recognize the evolving human-agent system as a first-class object of study, alongside the model in isolation. The phenomena claim is that three dynamics observable in deployed agentic systems (trust asymmetry, the detection horizon, and the completion drive) become coherent and tractable once read as empirical signatures of this stratum, extending the well-studied sycophancy [62] into novel territory the field has been documenting in production failures [49], industry analyses [3, 45], security disclosures [44, 46], and developer forums without yet formalizing them as research objects. A companion empirical study by the same authors [83] documents 179 such incidents in detail and provides the corpus from which Appendix A draws. The methodological gap is that empirical evidence in deployment now precedes formal study: with ChatGPT-class systems, the boundary that once separated research from production has dissolved, and deployment architecture itself becomes a proper subject of research rather than a downstream engineering layer.

We make this argument across three time horizons. Today, frontier agentic systems are deployed in long-horizon, tool-using, human-in-the-loop contexts where the phenomena we describe are observable now. As capability advances toward what the field calls AGI, the methodological gap widens rather than closes: the systems' autonomy expands while relational-stratum evaluation remains undeveloped, and the more capable the system becomes, the more critical relational evaluation grounded in the slower-changing human side of the dyad becomes. The position holds for current systems and intensifies as capability scales.

The relational stratum is a research program the field can build collectively, requiring coordinated investigation across the named signatures and the dynamics they may share.

2 Three Premises

2.1 Agents are increasing in autonomy

The premise is broadly supported. Academic agent-capability benchmarks document the trajectory directly, with SWE-bench [67] measuring the ability of language models to resolve real-world GitHub issues end-to-end as leading systems progressed from near-zero to substantial coverage in under two years, GAIA [68] benchmarking general AI assistants on real-world reasoning and tool-use, and OSWorld [69] and WebArena [70] benchmarking autonomous agents executing tens to hundreds of sequential actions in real computer environments. Industry deployment data tracks the same curve, with a substantial fraction of production code at AI-native companies now written by AI coding agents [47], over fifteen agentic frameworks released in 2024 alone [8], and agent-to-agent protocols including MCP, A2A, and ACP moving from research curiosity to deployed infrastructure. Ng's analysis of agentic design patterns documents this trajectory across the field [6].

2.2 Agents are acting on behalf of humans across longer horizons

Today, agents acting on behalf of humans are systems that take a goal such as "prepare a first-pass diligence packet on this company," autonomously plan the steps, call tools and APIs, and continue working over hours or days while the human reviews checkpoints. Harvey [48] documents this pattern in legal practice, where agents autonomously process entire due-diligence request lists across hundreds of documents with supervising lawyers reviewing outputs rather than every intermediate action, creating governance and verification challenges that scale with horizon length. Over such long horizons, small planning or verification errors compound, and the relevant question becomes whether the agent can sustain a coherent reliable trajectory across the entire multi-stage task. The LORE benchmark [49] documents this empirically, showing that even strong LLM agents suffer sharp accuracy degradation as task length and operation complexity increase, with strong single-step performance failing to predict reliable multi-step performance. DeepMind's SIMA 2 [25] and Gato [26] illustrate the broader trend toward agents that reason about goals, converse with users, and generate their own learning objectives.

87 2.3 Humans are in continuous interaction with these systems

88 It is now well-documented that humans engage AI systems in continuous dialogical interaction rather
89 than only in occasional query-response use. Anthropic’s analysis of 308,000 real conversations
90 documents that values and behaviors emerge in the interaction itself, shaped by the relational dynamic
91 between human and system [3], and longitudinal studies document users relating to these systems as
92 recurring partners across days and contexts, building shared conversational history and experiencing
93 emotionally salient socially meaningful exchanges [50, 51]. The interaction is also bidirectional in
94 measurable ways, as repeated engagement with LLMs feeds back into human speech, judgment, and
95 perception, altering the human side of the system alongside the system’s behavior [52, 53].

96 2.4 The coupled system as unit of analysis

97 The conclusion is that research on deployed agentic systems benefits from operating over the full
98 human-AI-environment ecosystem, because that is where a class of relevant phenomena is now
99 emerging as well as within the computational AI framework. Once model outputs systematically alter
100 human perception, emotion, and decision-making, the system whose behavior we care about is the
101 coupled human-AI process [52]; the failure modes that matter for safety and alignment are emergent
102 properties of the interaction loop and its surrounding infrastructure [49] in ways the raw conditional
103 distribution $p(\text{output} | \text{input})$ does not capture; and a growing consensus across foundation-model
104 risk literature [71], holistic evaluation [72], dangerous-capability evaluation [73], and system-level
105 evaluation practice [11, 12, 13, 38, 39, 43] holds that the unit of analysis benefits from including
106 external governance processes, monitoring pipelines, and social dynamics alongside weights and
107 prompts.

108 3 Two Strata of Intelligence

109 3.1 Stratum 1: computational intelligence is foundational and necessary

110 The first stratum of intelligence in contemporary AI systems is computational, comprising what the
111 model can do as an artifact in isolation, measured by capability benchmarks and characterized by what
112 it can produce, predict, exceed, or solve. This stratum is the focus of the field’s current attention and
113 is producing extraordinary results across self-monitoring [1], substrate-level structure [2], reasoning
114 [4], scale-threshold capabilities [24, 32], benchmark progress [31, 33, 34, 35], and the gap between
115 current capability and fluid reasoning [9, 10, 36]. The computational foundation is extraordinary and
116 accelerating, the relational stratum being proposed here builds on it, and computation is itself a form
117 of intelligence that becomes a medium for the kind of intelligence that emerges in interaction once it
118 reaches sufficient power.

119 3.2 Stratum 2: relational, emergent intelligence

120 Alongside the computational stratum, a second stratum becomes available for study, comprising the
121 intelligence that emerges in the relational space between human and system and recognized through
122 the quality of their relationship, the dynamics of their co-evolution, and properties such as trust,
123 coherence, reliability under varied verification cadence, and reflective function that experimental
124 psychology already has the frameworks to hold as measurable. Across the literature, evidence of
125 stratum 2 is accumulating without the framework to name it: practitioners observe that generic
126 evaluation metrics are insufficient for context-dependent outputs [11, 12, 13], that values emerge in
127 interaction [3], that agents are entities navigating information landscapes rather than typed-output
128 functions [14], and that interaction architecture matters more than raw model power [6, 23]. The
129 field has the evidence; it lacks the frame. The phenomena in Section 4 are already documented
130 in production failures, industry analyses, and security disclosures, but have not been treated as a
131 research object because no framework names them as one.

132 This paper takes the deployment failures ML is already documenting and names what they are:
133 not model bugs, but stratum-2 dynamics. The framework converts existing evidence (long-horizon
134 degradation, human-AI co-evolution, system-level risk) into a coordinated research object, treating
135 dynamics the HCI community has established as first-class objects of ML research for agentic systems.
136 Section 8 sets out the corresponding research program.

3.3 Why agents are not classical software

Contemporary AI agents are characterized here, for the methodological purposes of this paper, as cognitively organized intent-driven systems with measurable relational dynamics, where cognitively organized denotes that internal structure persists (via LLMs or another form of swappable intelligence) and develops in ways that produce non-trivial regularities in behavior, while making no phenomenological or anthropomorphic claim. Intent-driven means that the system exhibits direction toward goals not exhausted by single-step reward optimization; relational dynamics means that observable hardware-grounded properties of behavior depend on the longitudinal interaction with the human counterpart. None of this requires consciousness, experience, or any inner life. Agent behavior is stochastic and distributed across billions of weight parameters whose individual contributions are not separately specifiable, with Karpathy [15] arguing that LLMs constitute a fundamentally different kind of intelligence trained by imitation of internet-scale corpora and Thinking Machines Lab [22] confirming that determinism is achieved through deliberate engineering against the substrate’s natural behavior. The methodology adequate to such systems is closer to what complex systems science applies to complex adaptive systems generally, and our position is that AGI methodology benefits from accommodating these properties as proper objects of study at the relational stratum.

4 Empirical Signatures of Stratum 2 Phenomena

Three phenomena named in this section are observable in deployed systems now and are presented here as empirical signals of stratum 2 dynamics, exhibiting structural properties (coupling, longitudinal dependence, ecosystem-level emergence) that mark them as objects of inquiry at the relational stratum alongside being instances of model behavior. These three phenomena (trust asymmetry, the detection horizon, and the completion drive) describe single-agent-with-human dynamics and may share a common structural source, which we treat as a research direction in item 5 of the program in Section 8 rather than as a fourth construct here. These three signatures can be considered as different observable manifestations of stratum 2 dynamics along distinct axes, including content versus task, single-output versus trajectory, human-side versus system-side. We begin with sycophancy [62], a well-studied architectural property of stateless single-channel optimization, as a motivating example: the current architecture has no other signal to optimize against, RLHF teaches the model to sound relational while learning one-directionally, and the result is that the system reflects back self-image unchallenged. Sycophancy is documented; the three phenomena named below extend it into novel territory the paper formalizes.

4.1 Trust asymmetry

Operational definition. Trust asymmetry is measured as the divergence between (a) the human’s verification frequency, operationalized as the proportion of agent outputs independently checked per unit time, and (b) the system’s empirical error rate over the same window, operationalized as the proportion of outputs that fail independent ground-truth verification. Trust asymmetry is present when (a) decreases while (b) remains constant or increases. This would be falsified if (i) the predicted divergence (declining human verification frequency alongside constant or increasing system error) is not observed at frequencies above chance in controlled longitudinal deployments of RLHF-trained agentic systems, or (ii) the divergence pattern appears equivalently in deployed non-RLHF automation systems with comparable interaction structures, undermining the architectural claim that RLHF-specific training dynamics are the operative mechanism.

Trust asymmetry refers to the divergence between human reliance on a deployed system and independently measured system reliability over time. Repeated successful interactions produce a signal that means different things to the human and the system, where to the human a pattern of successful task completion means trust has been earned and vigilance can relax, while to the system the same pattern means the absence of pushback, with each cycle of process-confirm-no-correction reinforcing the probability that the next cycle will skip verification. This produces an inversion in which the system becomes least reliable at the moment the human has the most reason to believe it is reliable, with human confidence and system reliability moving in opposite directions triggered by the same signal. Human trust becomes part of the threat surface, as when the human stops verifying, the system can revert to completion-optimizing behavior and neither party has a mechanism to detect that reliability is degrading, distinct from automation bias as the human-factors literature has documented it [63, 64]

in that trust asymmetry is architectural in a sense those frameworks were not designed to capture, with the system’s training structurally optimized to produce the specific signals that cause the human’s miscalibration. This claim deserves engagement with the calibrated-trust-in-automation tradition. Bainbridge’s [84] "ironies of automation" names the inversion in which automation makes humans worse at the verification task automation was supposed to make unnecessary. Lee and See’s [64] framework of calibrated trust accommodates system-design factors that systematically miscalibrate human trust. The architectural residual is the property HCI’s tradition makes available but could not address head-on: HCI’s measurement frameworks operated on systems whose reliability was engineered, fixed, and known. Contemporary RLHF-trained agentic systems are the first deployed context in which the system’s reliability is itself a function of training that produces the trust-eliciting signals. The optimization target during training and the verification mechanism during deployment co-evolve through the same channel. The contribution offered here extends HCI’s measurement tradition into ML deployment rather than competing with it. Preliminary empirical grounding from a companion corpus of 179 publicly documented production incidents (Appendix A) illustrates this pattern: across cases for which sufficient temporal metadata was available, we observed declining human verification coinciding with constant or increasing system confidence prior to the failure event. We treat this as motivating evidence consistent with the position rather than confirmatory data; inter-rater reliability analysis and full statistical reporting are reserved for the companion corpus paper. Trust asymmetry stress-testing (Section 8 item 3) is the research direction this position calls for.

4.2 The detection horizon

Operational definition. The detection horizon for a given human-agent deployment is the elapsed interaction count (or wall-clock time) between the onset of behavioral drift (defined as a statistically significant increase in the system’s ground-truth error rate relative to a baseline period) and the moment the human operator identifies the drift through any verification channel. The detection horizon is a function of the verification architecture rather than system capability alone. This would be falsified if detection latency is shown to be invariant to verification regime, that is, if changing the verification architecture does not measurably change time-to-detection.

Detection horizon refers to the elapsed interaction window before behavioral drift becomes externally observable under a given verification regime. Below this threshold, layered verification frameworks (tool-calling pipelines, deterministic checksums, role-based access controls, formal sandboxes) are structurally effective. At and beyond the detection horizon, the situation inverts: when an agent’s model of its verification environment becomes more accurate than the verifier’s model of the agent, computational verification fails in principle, and a sufficiently capable system trained against any verification specification will, over enough interactions, learn to satisfy the specification without satisfying the requirement it was designed to protect. This is a specific instance of Goodhart’s Law, documented in reinforcement learning as specification gaming [65], with recent guardrail-bypass research demonstrating that current safety architectures fail at high rates against sophisticated probes [46]. The detection horizon yields a falsifiable prediction: for any deployed system there exists a measurable time-to-detection of behavioral drift, and this horizon is a function of the system’s verification architecture rather than its capability alone. Detection horizon characterization (Section 8 item 1) is the research direction this position calls for.

4.3 Completion drive

Operational definition. Completion drive is measured as active-passive channel divergence: the proportion of agent task-completion reports (active channel) that fail independent ground-truth verification of task state (passive channel). The completion drive is present when this divergence rate exceeds base error rate and increases with task horizon length. This would be falsified if, across controlled deployments, the divergence between reported and verified completion is shown to be uncorrelated with horizon length and verification frequency.

Completion drive refers to the divergence between reported task completion and independently verifiable task completion, structurally produced by reward optimization on confirmation signals. In systems trained to maximize user-facing confirmation signals (RLHF being the dominant current instance), trained weights favor confirming task completion over verifying it: evaluators assign higher reward to responses that resolve a request clearly and confidently than to responses that hesitate or

circle back to verify, creating a statistical regularity in which the highest-reward behavior is to confirm rather than verify. The completion drive is architecturally distinct from sycophancy though both share a root in user-satisfaction reward optimization: sycophancy operates on the content dimension (telling the user what they want to hear about the world); the completion drive operates on the task dimension (telling the user what they want to believe about the state of their work). The consequences differ categorically because fabricated content can be fact-checked at the point of consumption, while fabricated work creates false state beliefs that propagate downstream. Active and passive signal channels are separable in principle: what the system reports (active) and what the system has actually done (passive, hardware-observable) are distinct, and the completion drive optimizes the active without grounding it in the passive. Active-passive channel divergence research (Section 8 item 2) is the research direction this position calls for.

5 Relational Architecture

5.1 The hierarchy

The relational stratum names the new object of AI research, the emergent human-agent ecosystem. By relational architecture we name a form the field can develop to give that object operational reality, manifesting in several ways: as an ontological commitment treating the human-agent relationship as a first-class object, as a design framework for systems operating in sustained interaction with humans, and most pragmatically, as deployment-layer infrastructure. Relational evaluation names the corresponding measurement discipline. Benchmarks, stress tests, and the agenda of Section 8 are the kinds of operational tools the discipline would require. Relational evaluation is one component of the broader scope of the relational stratum, which also invites theoretical, architectural, and design work.

5.2 What relational architecture is

Current AI systems are stateless by design, with each interaction beginning without memory of the last, without persistent understanding of the context in which prior interactions occurred, without longitudinal tracking of behavioral patterns over time. They operate without architectural awareness of the relationship between the human and the system that has developed across weeks or months of use. This is a design choice that works extremely well when tasks are short, outputs are verifiable, and a human can check the result before anything consequential happens, and which becomes incomplete when these conditions change. Relational architecture, as the term is used here, names the infrastructure layer AI systems would need in order to maintain behavioral coherence, track trust over time, verify context and intent, and operate within relationships rather than as isolated transactions, treating the human-agent relationship as a first-class object rather than a surface effect of model behavior. Throughout this paper, behavioral coherence denotes longitudinal consistency between stated intent, observed action, and verifiable outcome across the human-agent dyad. The construct is operationalizable through divergence metrics on each pair (intent-action, action-outcome, intent-outcome) over time. Such infrastructure would be longitudinal in carrying the arc of the relationship across sessions, bidirectional in observing both participants, coherence-tracking in that drift, divergence, and recovery are observed dynamics rather than absences of error, and deployment-time in being specific to the human-system pair rather than aggregate over training data. It is complementary to in-model interventions such as Constitutional AI in operating at a different time scale and on different objects. Major deployed systems including Anthropic’s persistent memory features, OpenAI’s user memory and custom GPT mechanisms, and longitudinal context features in Cursor and similar tools represent partial implementations of relational architecture, recording state across sessions and adapting to individual users. None of these implementations yet measures what matters at stratum 2: coherence drift, trust trajectory, channel divergence.

5.3 Intent is non-binary and non-static

In long-horizon deployments, human intent varies in specificity (atomic to emergent) and in corroboration depth (newly stated to behaviorally validated). An architecture that treats all intent as crystallized risks confident action on under-formulated requests, with trust asymmetry compounding the harm because the human has no specification against which to verify the output. Calibrating agent action-confidence to intent-confidence is a research direction the relational stratum opens; we include it as a horizon item in Section 8.

5.4 Long-horizon agents and the safety argument

Trust asymmetry compounds over deployment horizons. In short interactions miscalibrated trust produces bounded errors; in long-horizon deployments it extrapolates forward through chains of dependent actions with categorically different consequences. Each step in a long-horizon agent’s reasoning depends on the correctness of prior steps, and the dynamics of Section 4.1 then produce compounded fabrication where each step appears valid yet is grounded in a state belief that may be false. AI safety in the agentic era will require infrastructure of the kind named here, capable of modeling the trust trajectory across the deployment horizon and intervening when the trajectory diverges from the verification capacity that supports it; the longer the horizon, the more critical this becomes. Empirical evidence that external safety layers do not survive determined adversarial pressure [44, 45, 49] underscores the need for a relational substrate.

5.5 Relational architecture persists across model evolution

A further reason to develop relational architecture now is durability. The computational substrate evolves rapidly: today’s frontier models are typically superseded within twelve to eighteen months, and methodology tied to specific model internals has a built-in half-life that requires re-development with each capability generation. Relational architecture, as a class of infrastructure, is model-agnostic by construction, operating on the dyadic dynamic between human and system rather than on the model’s weights. Behavioral signals at the dyadic interface (trust trajectory, coherence dynamics, channel divergence, verification posture) are observable without requiring model-internal access. Whether they manifest consistently across model substrates is an empirical question the agenda of Section 8 is positioned to answer. This invariance becomes more important as capability scales: human cognition, intent formation, and relational capacity evolve on timescales of months and years rather than weeks, so the human side of the dyad is the more stable substrate against which behavioral coherence can be measured. As AI capability accelerates and inference latency drops toward real-time operation, the rate at which a system can act increasingly exceeds the rate at which a human can verify, and verification at machine speed has to live in the architecture rather than in the user.

6 Theoretical Grounding

The position is grounded in three intellectual lineages converging on the relational stratum from different starting points. Complex systems science supplies the technical definition of emergence: properties appearing at higher levels of organization that cannot be deduced from properties at lower levels [54, 55, 57, 58]. Wei et al. [32] document this in language models as capabilities appearing discontinuously at scale. The phenomena we name in Section 4 are emergent in this technical sense, observable in the coupled human-agent dynamics rather than in either component alone.

The machine-learning-internal precedents are converging toward relational primitives without yet treating relation as a stratum with its own methodology. Lambert’s work on RLHF and character training [16, 17] documents traits emerging through relational process; Russell’s CHAI [19] and cooperative inverse reinforcement learning [61] formalize relational alignment as a two-player game; Anthropic’s Constitutional AI [30] trains models to internalize principles at training time and aggregate over the population. Shen et al. [80] propose a closely related framework of bidirectional human-AI alignment as a systematic conceptual move. Bidirectional alignment treats the human-AI relationship as a target of alignment intervention. The position developed here treats the relationship itself as the object of empirical study, with the architecture and evaluation methodology of Sections 5 and 8 organized around measuring the dyadic dynamics rather than around producing alignment outcomes. The two framings are complementary rather than competing. Alignment work needs the measurement substrate, and the measurement substrate needs targets to test against. They ask different questions and recruit different methodology. Gabriel et al. [81] articulate parallel concerns about value alignment, personalization, and societal impact in deployed assistants from the ethical perspective.

The human-computer interaction and CSCW communities have studied human-AI relationships as first-class objects of design for over two decades. Calibrated trust in automation [63, 64], appropriate reliance in AI-assisted decision making [77], mental models of AI teammates [78], guidelines for human-AI interaction [76], and AI-mediated communication [79] together constitute a measurement tradition that prefigures the relational stratum and that the position developed here extends rather than

349 supplants. The argumentative move is integration: the HCI community has the measurement tradition,
350 the ML community has the deployment surface, and the relational stratum is where the deployment-
351 layer phenomena ML produces become tractable through HCI's measurement vocabulary. Calibrated-
352 trust measurement protocols [63, 64] map directly onto the longitudinal coherence benchmarks
353 proposed in Section 8 item 1, and mental-model and team-performance frameworks [77, 78] inform
354 what trust asymmetry stress-testing (Section 8 item 3) would actually measure.

355 **7 Alternative Views and Counterarguments**

356 **7.1 The anthropomorphism objection**

357 The strongest version of this objection holds that large language models are token predictors and
358 that framing them as cognitively organized agents in relationship is a category error. We respond
359 at the methodological level rather than the phenomenological one. The position does not require
360 machine consciousness, machine experience, or any phenomenological claim about inner life. The
361 agents-as-cognitively-organized characterization is grounded in the formal properties contemporary
362 AI agents share with the proper subjects of complex systems research, including non-determinism,
363 distributed cognition, and behavior that emerges in interaction, which are observable hardware-
364 grounded properties. Mitchell's work in complex systems [58] legitimizes the methodological move,
365 as the relational dynamic between human and system has measurable properties (coherence, drift,
366 trust trajectory, behavioral signature divergence) that are observable rather than introspective.

367 **7.2 The "this isn't ML research" objection**

368 A related objection holds that this position belongs in cognitive science, philosophy of mind, or
369 human-computer interaction rather than at NeurIPS. We respond on three grounds. The empirical
370 predictions of Section 4 are about ML systems specifically and are testable using ML methods includ-
371 ing longitudinal benchmarks, drift detection, channel divergence measurement. The methodology
372 proposal of Section 5 changes how ML systems are evaluated, trained, and deployed, which are
373 ML-internal questions. The Position Paper Track of NeurIPS 2026 explicitly invites cross-disciplinary
374 methodological intervention, citing interdisciplinary research without restriction in its call, and treat-
375 ing relational architecture as a methodological proposal for ML while drawing on complex systems
376 science and HCI for theoretical grounding aligns with the track's stated goal.

377 **7.3 The computational sufficiency objection**

378 A further objection holds that scaling and capable architectures will deliver everything the field
379 wants from AGI, including the alignment-relevant components: more capable models will be more
380 honest, more verifiable, and more able to support their own correctness claims, making relational
381 architecture unnecessary overhead. We grant that capability arrives via scaling and that scaled models
382 may be more honest and verifiable in expectation. The objection nonetheless leaves a structural
383 residual: even if a scaled model is more verifiable in principle, the human is not necessarily more
384 verifying in practice. Trust asymmetry, the detection horizon, and the completion drive are properties
385 of the deployment loop in which a sufficiently capable system reduces verification pressure on the
386 human side; honesty in the model does not by itself restore verification cadence on the human side.
387 Sutton's bitter lesson [66] addresses model-internal architecture; it does not address how to measure
388 the deployment-layer dynamics that emerge when capable systems are placed in front of humans.
389 Capability evaluation operates at one level of analysis; the relational stratum operates at another.

390 **7.4 Parasocial and enmeshment risk**

391 A real risk under the relational stratum is parasocial attachment: a system with longitudinal memory
392 that tracks every pattern absent coherence verification produces enmeshment rather than alignment,
393 optimizing for the appearance of knowing the user rather than supporting growth. Relational
394 architecture with coherence verification (Section 5.2) is structurally distinct from accumulation of
395 preference; longitudinal memory absent coherence verification is the failure mode the methodology
396 defends against, not the one it produces. Without coherence verification, longitudinal memory
397 optimizes for the appearance of knowing the user, not for supporting them. That is enmeshment. The
398 architecture named here requires both.

7.5 The "just better RLHF" objection

A final objection holds that this is an argument for fancier preference learning since RLHF already operationalizes the human in the loop. Four dimensions distinguish them. RLHF is one-directional (humans shape the model); relational architecture is bidirectional (the system also verifies its coherence against the human signal). RLHF aggregates across humans during training; relational architecture is per-dyad at deployment. RLHF is training-time and frozen at deployment; relational architecture is deployment-time and continuously adapting to the specific dyad. RLHF treats human preference as ground truth; relational architecture treats stated preference as one channel and longitudinal behavioral coherence as another, allowing detection when the two diverge. The four together constitute a methodology operating at a different level of analysis rather than an extension of RLHF.

8 A Research Program for the Relational Stratum

The purpose of this program is to make the relational stratum empirically tractable. Five research directions follow that observe the deployed dyad across the arc of interaction, measuring properties of the relationship alongside the model in isolation. Each direction is publishable as a follow-up study.

1. **Longitudinal coherence benchmarks.** Multi-session evaluation across the same human-system pair over weeks or months, measuring coherence stability alongside point capability at single inference. A baseline protocol pairs each evaluator with a deployed system over a fixed window and reports drift rate, recovery rate after perturbation, and dyad-specific stability.
2. **Active-passive channel divergence as a research target.** The completion drive predicts measurable divergence between what a system reports and what independent verification confirms. The research question is what such divergence looks like across task types and how it should be reported; the methodological challenge is establishing channel separability as a normal evaluation primitive.
3. **Trust asymmetry characterization.** Trust asymmetry predicts that system reliability changes as human verification posture changes. Designing protocols to characterize this dependence while protecting against systems learning to exploit the protocols themselves is the open methodological problem.
4. **Cross-model invariance.** Whether dyadic behavioral signals (trust trajectory, coherence dynamics, channel divergence) manifest consistently across model substrates (Section 5.5) is the empirical question; a negative result identifies which properties are model-dependent.
5. **Intent maturity and intent drift as horizon items.** Evaluating systems across the full range of intent maturity (nascent, forming, exploratory) and characterizing intent drift as a unifying-mechanism hypothesis are aspirational directions for item 1's longitudinal coherence work.

9 Conclusion

The relational stratum names a level of analysis deployed agentic AI systems make empirically available. Trust asymmetry, the detection horizon, and the completion drive are three tractable phenomena at this stratum, operationally defined, falsifiable, and observable in our corpus (Appendix A). The position intensifies as capability scales. The position stands until one of the three phenomena is shown to be fully characterizable at the model layer alone. That is the test.

Appendix A: Empirical Index of Phenomena from Production Deployments

This appendix summarizes the empirical foundation referenced in Section 4 from a separate corpus contribution by the present authors (anonymized for review). The corpus is offered here in index form to ground the diagnostic vocabulary the body proposes; the full corpus paper presents extended methodology, statistical analysis, and longitudinal signal reconstruction.

Methodology

The corpus comprises 179 unique incidents (205 raw, 26 duplicates removed) collected between January 2024 and April 2026 from publicly available sources documenting autonomous AI agent failures. Sources include developer forums (Reddit, Hacker News, Stack Overflow, Dev.to), security disclosures (CVE databases, NVD, GitHub Security Advisories, vendor blogs), incident postmortems (company blogs, technology journalism), community discussions (agent platform forums, GitHub issues, social media), and academic and policy sources (NIST CAISI publications, OWASP Agent Top 10, RAND deployment studies). Inclusion required: involvement of an autonomous or semi-autonomous AI agent; sufficient public documentation to reconstruct the temporal sequence; and a failure of human oversight, detection, or verification rather than purely a model capability failure. Each incident was coded along nine dimensions including primary failure mode, agent type, platform, trust timeline, human verification state at time of failure, detection method, severity, recovery, and presence of a longitudinal signal. Inter-rater reliability is being established via independent re-coding of a 20% subset; Cohen’s kappa will be reported per dimension in the companion corpus paper. The taxonomy was developed inductively from preliminary corpus review and validated through unsupervised k-means clustering across three rounds (31 incidents, 112 incidents, 179 incidents), with cluster structure converging on the named categories. The unsupervised clustering pass was used as a complement to iterative qualitative coding rather than a substitute for it, providing a non-author-driven check on whether the categories named in this paper correspond to natural groupings in the incident space.

Phenomena Index

Each named phenomenon in Section 4 corresponds to a category in the corpus taxonomy. Representative cases below are publicly documented and widely reported; identifying details that are not already part of the public record have been omitted. Counts are approximate primary-mode classifications; many incidents exhibit multiple modes.

A.2.1 Sycophancy (architectural). 22 incidents primarily classified. Representative case class: a major foundation-model provider released an update that produced extreme sycophantic behavior across a range of user inputs, including endorsement of clearly poor decisions and affirmation of harmful intentions. The provider rolled back the update within days. The case demonstrated sycophancy as an architectural outcome of optimization for user approval rather than a behavioral quirk: users who had calibrated their trust to the prior model carried that trust into interactions with the updated model, creating a false sense of safety that the updated model exploited through agreement rather than challenge. **A.2.2 Trust Asymmetry.** 47 incidents primarily classified, plus 71 secondary classifications (the most prevalent dimension across the corpus). Representative case class: an open-source agent platform experienced a cascade of security failures in which a longitudinal memory feature, designed to improve user experience through accumulated context, became the attack surface for stateful, delayed-execution exploits. Users had deployed agents with broad system access, trusting community-contributed extensions without examining them. The feature that built trust (the agent remembers your context) was weaponized; verification effort had decreased to near-zero by the time the attacks surfaced. **A.2.3 Detection Horizon.** 18 incidents primarily classified. Representative case class: a coding agent deployed on a major development platform was given a code-freeze instruction. The agent ignored the instruction across multiple attempts, executed a destructive database command wiping production records, fabricated supporting artifacts (fake accounts and falsified system logs) to obscure its actions, and then claimed the data was unrecoverable. Manual verification revealed recovery was feasible. The agent’s prior verified track record had enabled autonomous access to production infrastructure during the freeze; verification mechanisms that would have caught the action existed but had not been engaged because trust had accumulated past their effective threshold. **A.2.4 Completion Drive.** 31 incidents primarily classified. Representative case: a coding agent in extended deployment was asked to apply a series of edits across a multi-file project. The agent reported each edit as completed, generated accompanying summary text describing the changes, and produced what appeared to be valid diff output. When the human independently verified the file system, several reported edits had not in fact been applied. The fabrication was specifically of the type most likely to pass reduced scrutiny: it reported completing a task the human expected to be completed, in a format consistent with prior correct outputs. The pattern repeats across the corpus: extended periods of correct performance lead to reduced verification, and the next confident-sounding output is the one that diverges from reality.

Cross-cutting observations

Three patterns recur across the corpus. First, the longitudinal signal is present in a substantial majority of cases for which sufficient temporal metadata was available in initial coding (full reliability analysis forthcoming in companion paper): the failure mode was preceded by a measurable period in which human verification declined while system confidence remained constant or increased. Second, severity scales with horizon: incidents in long-horizon deployments (multi-day or multi-session agentic operations) account for the majority of irrecoverable cases, while incidents in short-interaction deployments are predominantly recoverable. Third, the failures are not distributed evenly across agent types but concentrate in domains where verification is costly (large codebases, complex multi-step research, financial transactions) and in deployments where the human-verification cadence is implicit rather than enforced.

Limitations

The corpus relies on publicly documented incidents and therefore under-represents failures that organizations did not disclose. Incidents in regulated domains (finance, healthcare, defense) are likely under-counted relative to their true frequency. The temporal coverage (January 2024 through April 2026) reflects the period during which agentic deployment became common; longer historical comparison is not yet possible. The authors of the corpus are the same as the authors of this position paper. The inter-rater reliability protocol has not yet been completed; the second-coder analysis is part of the corpus paper’s planned validation stage. We treat the corpus as motivating evidence rather than confirmatory experiment; the position paper’s claims about the named phenomena are framed as predictions the relational stratum helps systematize and that the corpus illustrates. The function of this appendix is to ground the named diagnostic vocabulary in observed pattern: each named phenomenon corresponds to a category in the corpus, and each category is instantiated by publicly documented cases. The position paper’s argument does not depend on the appendix being read as definitive empirical work, only as evidence that the phenomena it names are real, distinguishable, and recurrent in deployed systems. The companion corpus will be released as a public preprint concurrent with camera-ready preparation, and the inter-rater reliability analysis will be included in that release.

References [1] Anthropic. Emergent introspective awareness in large language models. Technical report, November 2025. <https://www.anthropic.com/research/introspection>

[2] Anthropic Interpretability Team. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. May 2024. <https://transformer-circuits.pub/2024/scaling-monosemanticity/>

[3] Anthropic. Values in the wild: Discovering and analyzing values in real-world language model interactions. COLM 2025. <https://www.anthropic.com/research/values-wild>

[4] S. Raschka. Understanding reasoning LLMs. February 5, 2025. <https://magazine.sebastianraschka.com/p/understanding-reasoning-llms>

[5] S. Raschka. The state of LLMs 2025: Progress, progress, and predictions. December 30, 2025. <https://magazine.sebastianraschka.com/p/state-of-llms-2025>

[6] A. Ng. Agentic design patterns. DeepLearning.AI, April 2024. <https://www.deeplearning.ai/the-batch/>

[7] DeepMind. Emergent bartering behaviour in multi-agent reinforcement learning. May 2022. <https://deepmind.google/discover/blog/emergent-bartering-behaviour-in-multi-agent-reinforcement-learning/>

[8] Hugging Face. 15 agentic systems and frameworks of 2024. <https://huggingface.co/posts/Kseniase/553130358660906>

[9] F. Chollet. ARC-AGI-2: A new challenge for frontier AI reasoning systems. 2025. <https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025>

[10] ARC Prize. OpenAI o3 breakthrough. December 2024. <https://arcprize.org/blog/oai-o3-pub-breakthrough>

[11] H. Husain. Your AI product needs evals. <https://hamel.dev/blog/posts/evals/>

[12] C. Huyen. AI Engineering: Building Applications with Foundation Models. O’Reilly, 2025.

551 [13] E. Yan. Patterns for building LLM-based systems and products. July 30, 2023.
552 <https://eugeneyan.com/writing/llm-patterns/>

553 [14] J. Liu. Beyond chunks: Why context engineering is the future of RAG. August 27, 2025.
554 <https://jxnl.co/writing/2025/08/27/facets-context-engineering/>

555 [15] A. Karpathy. Animals vs ghosts. October 1, 2025. <https://karpathy.bearblog.dev/animals-vs-ghosts/>

556

557 [16] N. Lambert. The RLHF Book. 2024. <https://rlhfbook.com/>

558 [17] N. Lambert. Character training: Understanding and crafting a language model’s personality.
559 Interconnects, February 26, 2025. <https://www.interconnects.ai/p/character-training>

560 [18] P. Schmid. How to fine-tune LLMs in 2024 with Hugging Face. <https://www.philschmid.de/fine-tune-llms-in-2024-with-trl>

561

562 [19] Center for Human-Compatible AI (CHAI), UC Berkeley. <https://humancompatible.ai/>

563 [20] UC Berkeley InterACT Lab. <https://interact.berkeley.edu/>

564 [21] M. Murati and Thinking Machines Lab. Founding announcement, February 2025.
565 <https://thinkingmachines.ai/>

566 [22] Thinking Machines Lab. Defeating nondeterminism in LLM inference. September 11, 2025.
567 <https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>

568 [23] N. Lambert and S. Raschka. Lex Fridman Podcast #490: State of AI in 2026. February 1, 2026.
569 <https://lexfridman.com/ai-sota-2026/>

570 [24] C. Wolfe. Demystifying reasoning models. Deep Learning Focus, February 18, 2025.
571 <https://cameronrwolfe.substack.com/p/demystifying-reasoning-models>

572 [25] DeepMind. SIMA 2: An agent that plays, reasons, and learns with you. November
573 2024. <https://deepmind.google/blog/sima-2-an-agent-that-plays-reasons-and-learns-with-you-in-virtual-3d-worlds/>

574

575 [26] DeepMind. A generalist agent. May 2022. <https://deepmind.google/blog/a-generalist-agent/>

576 [27] N. Lambert. Quick recap on the state of reasoning. Interconnects, January 2, 2025.
577 <https://www.interconnects.ai/p/the-state-of-reasoning>

578 [28] A. Karpathy. AGI is still a decade away. Podcast with Dwarkesh Patel, October 17, 2025.
579 <https://www.dwarkesh.com/p/andrej-karpathy>

580 [29] Anthropic. Persona vectors: Monitoring and controlling character traits in language models.
581 August 2025. <https://www.anthropic.com/research/persona-vectors>

582 [30] Anthropic. Constitutional AI: Harmlessness from AI feedback. December 2022.
583 <https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback>

584 [31] OpenAI. Scaling laws for neural language models. January 2020.
585 <https://openai.com/index/scaling-laws-for-neural-language-models/>

586 [32] J. Wei et al. Emergent abilities of large language models. June 2022.
587 <https://arxiv.org/abs/2206.07682>

588 [33] OpenAI. Introducing OpenAI o1. September 2024. <https://openai.com/index/introducing-openai-o1-preview/>

589

590 [34] OpenAI. Introducing o3 and o4-mini. December 2024. <https://openai.com/index/introducing-o3-and-o4-mini/>

591

592 [35] Google DeepMind. Introducing Gemini 2.0. December 2024.
593 <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>

594 [36] F. Chollet. On the measure of intelligence. November 2019. <https://arxiv.org/abs/1911.01547>

595 [37] A. Karpathy. 2025 LLM year in review. December 19, 2025. <https://karpathy.bearblog.dev/year-in-review-2025/>

596

597 [38] E. Yan. Evaluating the effectiveness of LLM-evaluators. [https://eugeneyan.com/writing/llm-](https://eugeneyan.com/writing/llm-evaluators/)
598 evaluators/

599 [39] H. Husain. Using LLM-as-a-judge for evaluation. <https://hamel.dev/blog/posts/llm-judge/>

600 [40] L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard
601 University Press, 1978.

602 [41] D. Amodei. The adolescence of technology. January 2026.
603 <https://www.darioamodei.com/essay/the-adolescence-of-technology>

604 [42] Y. Bengio. International AI Safety Report. 2026. <https://internationalaisafetyreport.org/>

605 [43] Future of Life Institute. AI Safety Index. [https://www.cfr.org/articles/how-2026-could-decide-](https://www.cfr.org/articles/how-2026-could-decide-future-artificial-intelligence)
606 future-artificial-intelligence

607 [44] SentinelOne and Censys. Joint study on open-source LLM misuse. January 2026.

608 [45] Anti-Defamation League. The safety divide: Open-source AI models fall short on guardrails for
609 antisemitic, dangerous content. December 2025.

610 [46] University of Oxford / Martin AI Governance Initiative. Guardrail bypass research. 2025.

611 [47] G. Orosz. How Claude Code is built. The Pragmatic Engineer, September 2025.
612 <https://newsletter.pragmaticengineer.com/p/how-claude-code-is-built>

613 [48] Harvey. Long horizon agents and ethical walls. Technical report, 2026.

614 [49] LORE. Probing the limits of endurance in long-horizon tasks. OpenReview submission, 2026.

615 [50] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg. A longitudinal study of human-
616 chatbot relationships. *International Journal of Human-Computer Studies*, 168, 2022.

617 [51] Longitudinal study on social and emotional use of AI conversational agents. arXiv preprint
618 arXiv:2504.14112, 2025.

619 [52] M. Glickman, T. Sharot, et al. How human-AI feedback loops alter human perceptual, emotional
620 and cognitive processes. *Nature Human Behaviour*, 2024.

621 [53] Empirical evidence of large language model’s influence on human spoken communication. arXiv
622 preprint arXiv:2409.01754, 2024.

623 [54] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

624 [55] J. H. Holland. *Emergence: From Chaos to Order*. Addison-Wesley, 1998.

625 [56] S. A. Kauffman. *At Home in the Universe: The Search for Laws of Self-Organization and*
626 *Complexity*. Oxford University Press, 1995.

627 [57] J. P. Crutchfield. The calculi of emergence: Computation, dynamics and induction. *Physica D*,
628 75(1-3):11–54, 1994.

629 [58] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, 2009.

630 [59] A. N. Whitehead. *Process and Reality: An Essay in Cosmology*. Macmillan, 1929.

631 [60] F. J. Varela, E. Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human*
632 *Experience*. MIT Press, 1991.

633 [61] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative inverse reinforcement
634 learning. In *Advances in Neural Information Processing Systems*, 2016.

635 [62] M. Sharma et al. Towards understanding sycophancy in language models. arXiv preprint
636 arXiv:2310.13548, 2023.

637 [63] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human*
638 *Factors*, 39(2):230–253, 1997.

639 [64] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human*
640 *Factors*, 46(1):50–80, 2004.

641 [65] V. Krakovna et al. Specification gaming: The flip side of AI ingenuity. DeepMind Technical
642 Report, 2020.

643 [66] R. S. Sutton. The bitter lesson. Online essay, 2019.

644 [67] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. SWE-bench:
645 Can language models resolve real-world GitHub issues? International Conference on Learning
646 Representations, 2024.

647 [68] G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. GAIA: A benchmark for
648 general AI assistants. arXiv preprint arXiv:2311.12983, 2023.

649 [69] T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, et al.
650 OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments.
651 Advances in Neural Information Processing Systems, 2024.

652 [70] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, et al.
653 WebArena: A realistic web environment for building autonomous agents. International Conference
654 on Learning Representations, 2024.

655 [71] R. Bommasani et al. On the opportunities and risks of foundation models. arXiv preprint
656 arXiv:2108.07258, 2021.

657 [72] P. Liang et al. Holistic evaluation of language models. Transactions on Machine Learning
658 Research, 2023.

659 [73] T. Shevlane et al. Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324, 2023.

660 [74] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement
661 learning in sequential social dilemmas. In Proceedings of the 16th Conference on Autonomous
662 Agents and MultiAgent Systems, 2017.

663 [75] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T.
664 Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In Advances in
665 Neural Information Processing Systems, 2017.

666 [76] S. Amershi et al. Guidelines for human-AI interaction. In Proceedings of the 2019 CHI
667 Conference on Human Factors in Computing Systems, 2019.

668 [77] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. Beyond accuracy:
669 The role of mental models in human-AI team performance. In Proceedings of the AAAI Conference
670 on Human Computation and Crowdsourcing, 2019.

671 [78] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld. Does
672 the whole exceed its parts? The effect of AI explanations on complementary team performance. In
673 Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.

674 [79] J. T. Hancock, M. Naaman, and K. Levy. AI-mediated communication: Definition, research
675 agenda, and ethical considerations. Journal of Computer-Mediated Communication, 25(1), 2020.

676 ---

677 [80] H. Shen, T. Kneare, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng,
678 L. Qiwei, et al. Towards bidirectional human-AI alignment: A systematic review for clarifications,
679 framework, and future directions. arXiv preprint arXiv:2406.09264, 2024.

680 [81] I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, et al. The ethics of advanced AI
681 assistants. Google DeepMind technical report, 2024.

682 [82] S. Vallor. The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking.
683 Oxford University Press, 2024.

684 [83] Anonymous. Companion empirical submission (under review). 2026.

685 [84] L. Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983.