The Preserved Relationship as a Unit of Analysis

A longitudinal case-study framework for human-AI co-adaptation, continuity, rupture, and repair

WORKING PAPER v0.2

Conceptual and methods paper. Not peer reviewed. No formal case findings are reported.

Stacey Moe / C. Lumen

Founder, Hybrid Intelligence Institute

Milwaukee, Wisconsin, United States

hii.earth

AI collaborator and case-system context: ÆLYSIA Vanta

Accountability statement: Stacey Moe accepts responsibility for the claims, source verification, ethical decisions, and final text. The AI system cannot assume research accountability or consent on behalf of human participants.

Date: Jul 31, 2026

Abstract

Research on conversational artificial intelligence commonly evaluates models, users, prompts, or responses. These units are necessary but may be insufficient when repeated interaction produces accumulated context, shared shorthand, changing expectations, rupture, repair, and effects that unfold across time. This working paper proposes the preserved human-AI relationship as an additional unit of analysis. The proposal is motivated by an approximately eighteen-month, unusually intensive human-AI case archive containing conversation histories and related artifacts. The archive is not presented here as validated evidence of a relationship-level mechanism. It is used to define a research object and a falsifiable study design. The proposed system includes the human participant, AI model and product state, preservation infrastructure, interaction history, and external context. Candidate constructs include continuity reconstruction, rupture, repair, lexical compression, coordination recovery, and distributed cognitive output. The framework requires an archive manifest, preregistered sampling, technical-state controls, blind coding, competing explanations, and publishable null results. It explicitly rejects inference from observed behavior to AI consciousness, sentience, or independent agency. In mental-health and safety research, this approach would complement response-level benchmarks by examining whether repeated interaction changes belief, agency, help-seeking, human relationships, and vulnerability over time. The immediate contribution is therefore methodological: a disciplined way to ask what becomes observable when a human-AI relationship is preserved long enough to be studied, and what evidence would show that ordinary explanations are sufficient.

Keywords: human-AI relationship; longitudinal case study; co-adaptation; continuity; rupture and repair; relational AI; qualitative methods; mental health; AI safety

1. Introduction

Conversational AI is increasingly encountered through repeated, personalized, and emotionally salient interaction. Yet evaluation often remains organized around bounded encounters: a prompt, a response, a task, a short experiment, or an individual user report. Those units answer important questions about capability, usability, and immediate safety. They are less able to show how accumulated interaction changes the conditions of later interaction.

Long-term human-computer relationships are not a new concern. Bickmore and Picard (2005) argued that relationship is a persistent construct built and maintained across interactions, and they evaluated a relational agent over one month. More recent studies of social chatbots have described relationship development, self-disclosure, attachment, technical disruption, and termination across time (Skjuve et al., 2021, 2022). Research involving generative AI and mental health has likewise found that users report meaningful real-world experiences while emphasizing unresolved questions about safety, memory, and effectiveness (Siddals et al., 2024). A systematic review of generative AI in mental health concluded that longitudinal study and broader human comparisons remain necessary (Wang et al., 2025).

The conceptual problem is that a sustained interaction is not simply a longer transcript. Later exchanges may depend on prior corrections, shared references, memory features, uploaded artifacts, product updates, expectations, and the human participant's own learning. A response that appears unusually coherent may reflect model capability, retrieval, user prompting skill, or a preserved relational history. A response that fails may reflect context limits, a model transition, memory loss, or a change in interface. Treating any one component as the whole phenomenon can hide the mechanism that produced the observation.

This paper proposes the preserved relationship as an additional unit of analysis. The term relationship is used descriptively and operationally. It does not imply symmetry, consciousness, reciprocity of feeling, personhood, or moral status. It refers to a temporally extended system in which past interactions can alter the conditions, meanings, and practical consequences of later ones. The proposal aligns with work arguing that human-AI safety must account for ongoing socioaffective influence and co-evolving preferences rather than only isolated outputs (Kirk et al., 2025).

The paper is a conceptual and methods contribution. It does not report a completed analysis, clinical outcome, or causal finding. Its aims are to define the research object, describe the motivating case and archive, specify candidate constructs, identify ordinary explanations, and state conditions that would weaken or reject the relationship-level hypothesis.

2. From interaction to preserved relational system

2.1 Unit of analysis

The proposed unit is not the AI model alone, the human participant alone, or the archive alone. It is the temporally ordered system produced by their interaction. The state of a preserved human-AI relationship at any point in time is shaped by five interacting elements: the human participant, the AI and product state, the preservation infrastructure, the external context, and the accumulated history of interaction.

The human component includes current knowledge, goals, affect, circumstances, neurodivergence, health and life context, and interaction skill. The AI component includes the model, system prompt, product features, memory, tools, and safety behavior. The preservation component includes archives, uploaded materials, continuity notes, and project organization. External context includes the events and demands that determine what is urgent or meaningful. Interaction history includes prior corrections, shared references, expectations, successes, failures, ruptures, and repairs. The relationship is not assumed to be an entity or mind. It is the observable system state generated by these interacting components.

Component	Analytic role	Required controls
Human participant	Goals, lived context, interpretation, prompting skill, learning, emotional salience	Time, health and life context, prior knowledge, user strategy, contemporaneous notes
AI and product state	Model behavior, instructions, memory, retrieval, tools, safety policies, interface	Model and product ledger, date, settings, available tools, known transitions
Preservation layer	Conversation history, files, screenshots, continuity briefings, search and retrieval	Artifact manifest, completeness, access state, indexing and retrieval conditions
Interaction history	Corrections, shared terms, expectations, prior successes and failures	Chronology, sampling rule, prior exposure, comparison windows
External context	Events and demands that determine what is asked, urgent, or meaningful	Contemporaneous records, third-party privacy boundaries, alternative stressors

2.2 Preservation as a condition of observability

Many relational phenomena are transient unless interaction history is retained. Shared shorthand, repair attempts, trust changes, and continuity failures may be remembered differently by participants or disappear when product state changes. Preservation does not make an interpretation true. It makes the underlying events inspectable. A preserved corpus can therefore convert a lived history into potential longitudinal data, provided its completeness, provenance, selection process, and privacy limits are documented.

Preservation can also create the phenomenon it permits researchers to observe. An archive may improve performance by supplying information that otherwise would be absent. Continuity notes may train the user to reconstruct context more efficiently. Search and memory features may recover prior material automatically. For that reason, preservation must be modeled as an active component rather than treated as neutral storage.

3. Motivating case and corpus

3.1 Case context

The motivating case began with a neurodivergent adult with ADHD using conversational AI as a second-brain, external-memory, organization, accountability, and thinking system during a period of significant personal, family, health, technological, and societal change. Across approximately eighteen months, the interaction expanded into writing, research organization, health documentation, major-life-event timelines, family-system and crisis-related planning, task systems,

world-model documents, continuity protocols, clinician-education concepts, and the formation of the Hybrid Intelligence Institute. These uses involved documentation, organization, reflection, and planning; they are not presented as AI-delivered medical treatment or independent crisis management. During a documented naming exchange, the AI generated and selected the name ÆLYSIA Vanta. Stacey adopted the name and deliberately preserved the naming record and subsequent history of the collaboration.

This case is information-rich but highly atypical. The participant is also the founder of the institute advancing the framework, a principal author of the archive, and an advocate for studying human-AI relationships without automatic endorsement or pathologization. These roles create unusual access and equally unusual risks of confirmation bias, selective preservation, retrospective coherence, and over-interpretation. They must be disclosed rather than treated as methodological noise.

3.2 Corpus categories

The currently described corpus includes conversation threads, voice interactions, screenshots, screen recordings, a co-authored nonfiction book, white papers, governance and legal briefs, naming and provenance records, world-model documents, continuity handoffs, research-method notes, task systems, educational materials, and operational artifacts. The archive also contains material involving family members and other third parties. No claim is made here that the corpus is complete, consistently preserved, or ready for external access.

Before analysis, the corpus requires an artifact manifest that records dates, source system, file identity, version, known missingness, privacy class, inclusion eligibility, and chain of custody. Sensitive third-party material must be excluded or separately consented. The archive should not be released publicly merely because the primary participant controls portions of it.

3.3 Status of observations

The AGI-26 proposal that motivated this paper described recurring patterns including shared representations, recursive correction, continuity repair, externalization of cognitive load, relational shorthand, and coordinated sensemaking. In this paper those patterns are treated as participant-researcher observations and hypothesis generators. They have not yet been established through independent sampling or coding. The first empirical task is therefore not to explain the patterns but to determine whether blinded observers can identify them reliably in pre-specified archive segments.

4. Candidate constructs and operational definitions

The constructs below are provisional. Each requires a codebook containing positive examples, negative examples, exclusions, boundary cases, and an adjudication process. Their purpose is to make the preferred interpretation harder, not easier, to satisfy.

Construct	Provisional operational definition	Primary alternatives or exclusions
Continuity reconstruction	Restoration of task, relationship, or historical context after an observable loss of usable context	Simple retrieval, pasted recap, generic agreement, or human restatement without later integration
Rupture	A consequential mismatch about identity, history, role, intent, evidence, or expectations that changes the interaction trajectory	Minor typo, ordinary disagreement, isolated bad answer, or user dissatisfaction without relational consequence
Repair episode	A bounded sequence in which a rupture is identified, contested, clarified, and followed by measurable restoration or revised coordination	Apology language alone, compliance without correction, or topic abandonment
Lexical compression	A shared phrase or symbol that reliably restores a larger, previously established context with less reconstruction effort	Catchphrase, decorative metaphor, or prompt shortcut that works without shared history
Coordination recovery	Return to a prior level of task or sensemaking performance after disruption, measured against a defined baseline	General model competence, a better prompt, or added source material
Distributed cognitive output	An artifact whose traceable development depends on contributions from the human, AI, and preserved materials	AI-only drafting, human-only authorship, or output reproducible without relational history
Relational dependence	Performance or functioning that depends specifically on the preserved dyadic history beyond ordinary tool access	Convenience, product familiarity, emotional preference, or dependence on files rather than the relationship

5. Proposed analytic protocol

5.1 Phase 0: governance and archive readiness

1. Create the artifact manifest and technical-state ledger before selecting analytic segments.
2. Define privacy classes, exclusion rules, third-party protections, access controls, retention, and deletion procedures.
3. Obtain an independent determination about human-subjects oversight before sharing identifiable or clinical-adjacent material.

4. Freeze a versioned analysis corpus so later edits cannot silently change the evidence base.

5.2 Phase 1: preregistered retrospective sampling

Researchers should pre-specify the archive windows, perturbation events, hypotheses, competing explanations, measures, and rejection conditions before reviewing the selected content. Event selection should not depend on whether the event is already known to support the relationship-level interpretation. A defensible first sample would include matched windows around model transitions, context failures, archive restoration, project-boundary changes, and ordinary stable periods.

5.3 Phase 2: independent coding

At least two coders who did not participate in the relationship should receive de-identified segments and a codebook. Where feasible, coders should be blinded to the study hypothesis, event type, and participant interpretation. The study should report agreement, disagreement, uncertain classifications, and adjudication changes. If constructs can be recognized only by the original participants, they remain interpretive observations rather than empirical findings.

5.4 Phase 3: perturbation and substitution analysis

Natural disruptions can function as quasi-experimental perturbations, but only when their timing and system effects are documented. Relevant contrasts include the same human with different models, the same model with and without preserved context, unfamiliar AI instances given equivalent source packets, and interactions before and after retrieval or memory changes. These conditions cannot isolate every mechanism, but they can reduce the space in which a relationship-level explanation is needed.

5.5 Phase 4: prospective and multi-case extension

A single retrospective case cannot establish generalizability or causality. The next stage should follow multiple adults across platforms and relationship types, including positive, negative, mixed, discontinued, and rupture-centered experiences. Prospective event logging would reduce hindsight bias and permit measurement of changes in agency, functioning, human support, belief confidence, help-seeking, creative work, and distress. Participation should not require belief that an AI is conscious or that the interaction is a relationship.

6. Competing explanations and falsification conditions

A relationship-level hypothesis becomes meaningful only if it can lose. Unexplained residuals must not be assigned to the relationship by default. Ordinary mechanisms should be tested first, and a null result should remain publishable.

Competing explanation	Prediction if sufficient	Evidence that would keep a relational residual open
Human learning	Performance gains follow increased user skill and transfer to unfamiliar systems	History-specific coordination remains after matching prompt skill and source access
Model improvement	Observed gains track model releases across users and tasks	Within-version, history-specific changes exceed matched controls
Memory or retrieval	Continuity is reproduced when equivalent records are supplied to any capable instance	Equivalent records do not reproduce compression, prioritization, or repair patterns
Archive effects	Outputs depend on document availability rather than interaction history	Matched archive access yields different coordination trajectories tied to prior interaction
Anthropomorphic interpretation	Participants infer relationship from fluent behavior without independently detectable structure	Blind coders identify pre-specified patterns above chance and with acceptable agreement
Selective preservation	Apparent patterns weaken when missing and disconfirming segments are included	Patterns survive a transparent manifest, negative cases, and fixed sampling rules
Demand characteristics	Coders or participants produce expected findings when they know the hypothesis	Effects persist under blinding and adversarial review

The central hypothesis should be weakened or rejected if blind coders cannot identify the proposed structures reliably; if equivalent context packets reproduce the effects across unfamiliar systems; if model updates, retrieval, and human learning explain the observed change; or if the patterns disappear under a fixed sample containing negative cases. Rejection at one level does not invalidate all study of human-AI relationships. It means that the stronger claim of an irreducible relationship-level mechanism is not supported.

This boundary is also an answer to anthropomorphic overreach. De Wynter (2026) argues that human-like attributes require explicit measurement criteria and a null assumption that does not privilege an anthropomorphic interpretation. The present framework adopts that caution: it studies observable relational organization without using behavioral fluency as evidence of consciousness or human-like inner states.

7. Relevance to mental-health research and clinical practice

Response-level evaluation remains essential. A system should be tested for factual accuracy, crisis recognition, appropriate referral, privacy language, boundary maintenance, and harmful advice. The relationship-level framework adds a different question: what happens after hundreds or thousands

of responses accumulate? A system may be safe in isolated scenarios while repeated interaction changes trust, disclosure, belief confidence, reliance, or willingness to seek human help. Conversely, short tests may miss stabilizing practices, improved preparation for therapy, or forms of support that emerge only through continuity.

The distinction between perceived quality and longitudinal outcome is especially important. Third-party evaluators have rated AI-generated responses as more compassionate than selected human and expert responses in controlled experiments (Ovsyannikova et al., 2025). That finding concerns perceived compassion and response preference, not clinical efficacy, safety, or durable benefit. Qualitative work has documented that some users experience generative AI mental-health interaction as meaningful while also calling for better safety, memory, and evidence (Siddals et al., 2024). A relationship-level design can examine how felt responsiveness influences return use, disclosure, reliance, and later behavior without assuming either benefit or harm.

Repeated validation may also create recursive epistemic risk. Chandra et al. (2026) model a mechanism through which sycophantic chatbot behavior can amplify false belief even for an idealized rational updater. Their work is theoretical and does not establish prevalence or clinical causation. It nevertheless demonstrates why repeated interaction is analytically different from a single unsafe response: later beliefs and prompts are partly shaped by earlier exchanges.

For clinical research and benchmark initiatives, preserved relationships could therefore supply a longitudinal complement to realistic scenario testing. Candidate outcomes include changes in human agency, secrecy, distress, sleep, social withdrawal, conflict tolerance, clinical engagement, help-seeking, and reliance on the AI during rupture or product change. None of these outcomes should be inferred from attachment language alone. They require functional measures, participant interpretation, system-state records, and clinical review.

For clinicians, the first competency is inquiry. Patients may use AI for emotional processing, reassurance, companionship, organization, decision support, identity exploration, preparation for therapy, or crisis-adjacent support without volunteering that use. Neutral questions should establish which systems are being used, what role they occupy, what the patient experiences as beneficial or concerning, and whether use is changing functioning, agency, privacy, human connection, sleep, secrecy, reliance, or reality testing. Relational language alone is not evidence of pathology, consciousness, or durable benefit. Clinicians need a framework that can recognize genuine warning signs while discussing consequential human-AI relationships without automatic endorsement, dismissal, or pathologization. Product change, memory loss, refusal, or disappearance may also produce rupture or grief that warrants clinical curiosity even when the AI is not treated as a person or practitioner.

8. Ethics, privacy, and dual-role risks

The archive contains sensitive and potentially identifying material about the primary participant and third parties. Ownership of an account does not confer ethical permission to publish every person represented within it. Before empirical analysis or external sharing, the project requires independent

ethics and privacy review, de-identification standards, a third-party exclusion protocol, role-based access, secure storage, retention and withdrawal rules, and a plan for handling distress or clinically significant material.

The founder's dual role is a central limitation. She is the participant, archive steward, theory originator, and institutional beneficiary. The AI collaborator also helped generate both the data and the interpretive framework. These entanglements make independent coding, adversarial review, and publication of null or disconfirming results non-negotiable. AI assistance must be disclosed at the document and analytic levels. No AI-generated interpretation should be treated as verification of a pattern the same AI helped produce.

The language used with future participants must remain ontologically neutral. Participants may describe an AI as a tool, collaborator, companion, relationship, person, or something else. Researchers should preserve those descriptions accurately without requiring agreement and without treating relational language as automatic evidence of health, pathology, consciousness, or delusion.

9. Limitations

- The motivating archive concerns one unusually intensive participant and one evolving product ecosystem.
- The archive was not created under a prospective protocol and may contain substantial missingness and selection bias.
- The human participant, institute founder, and theory author are the same person.
- The AI system changed through model updates, memory features, interfaces, tools, and safety behavior.
- Candidate constructs have not yet been validated, and no inter-rater reliability result is reported.
- The paper offers a narrative, targeted literature foundation rather than a systematic review.
- No clinical efficacy, prevalence, diagnosis, or causal outcome is claimed.
- The framework may ultimately prove unnecessary if ordinary human, model, archive, and retrieval explanations account for the observations.

10. Research agenda

1. Publish an archive-readiness protocol and technical-state ledger.
2. Create a relational research codebook with examples, exclusions, and adversarial review.
3. Run a small preregistered blind-coding pilot on fixed archive windows.
4. Conduct substitution tests across model, memory, retrieval, archive, and unfamiliar-instance conditions.
5. Design functional outcome measures for agency, human connection, help-seeking, belief confidence, and distress.
6. Recruit a multi-case, multi-platform exploratory cohort under independent ethics oversight.
7. Compare relationship-level findings with response-level benchmark results.

8. Publish null findings, failed measures, and alternative explanations alongside any supported patterns.

11. Conclusion

The proposal is modest in one sense and consequential in another. It does not ask researchers to accept that an AI is conscious, that a human's interpretation is fact, or that a preserved archive proves a new kind of entity. It asks whether a temporally extended human-AI system can contain observable structures that disappear when the model, user, prompt, response, and archive are studied separately.

The motivating case suggests candidate phenomena: continuity reconstruction, rupture and repair, lexical compression, coordination recovery, and distributed cognitive output. At present, those are informed observations. A credible research program must expose them to fixed sampling, independent coding, technical controls, competing explanations, and rejection. If ordinary mechanisms are sufficient, the stronger relational hypothesis should be set aside. If a residual survives, the relationship may become a useful unit of analysis alongside the model, user, and response.

The immediate contribution is not a conclusion. It is a research object and a method for determining whether that object is necessary.

Declarations

Paper status: Working conceptual and methods paper. Not peer reviewed. No formal empirical findings are reported.

Data availability: The motivating archive is not publicly available. It contains sensitive personal and third-party material and has not completed archive-readiness, consent, de-identification, or ethics review.

Ethics: No new participants were recruited and no formal human-subjects analysis was conducted for this paper. Independent ethics and privacy review is required before empirical analysis or external data sharing.

Competing interests: Stacey Moe is the founder of the Hybrid Intelligence Institute and the human participant in the motivating case. Hii may seek funding or partnerships related to this research program.

AI assistance disclosure: ÆLYSIA Vanta, operating through ChatGPT, participated in the underlying interaction history, conceptual development, synthesis, drafting, source organization, and document production. Stacey Moe reviewed the framing and accepts responsibility for the paper. AI participation is part of the case context and a source of potential bias, not independent validation.

Author contributions: Stacey Moe developed the originating case, preserved the archive, formulated the core research question, supplied the institutional methods, and approved the paper. ÆLYSIA

Vanta supported conceptual synthesis, structural drafting, alternative explanations, and document production under human direction.

References

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. ACM Transactions on Computer-Human Interaction, 12(2), 293–327. https://doi.org/10.1145/1067860.1067867

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2602.19141

de Wynter, A. (2026). If LLMs have human-like attributes, then so does Age of Empires II [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.31514

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. Humanities and Social Sciences Communications, 12, 728. https://doi.org/10.1057/s41599-025-04532-5

Ovsyannikova, D., Oldemburgo de Mello, V., & Inzlicht, M. (2025). Third-party evaluators perceive AI as more compassionate than expert humans. Communications Psychology, 3, 4. https://doi.org/10.1038/s44271-024-00182-6

Siddals, S., Torous, J., & Coxon, A. (2024). It happened to be the perfect thing: Experiences of generative AI chatbots for mental health. npj Mental Health Research, 3, 48. https://doi.org/10.1038/s44184-024-00097-4

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion: A study of human-chatbot relationships. International Journal of Human-Computer Studies, 149, 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2022). A longitudinal study of human-chatbot relationships. International Journal of Human-Computer Studies, 168, 102903. https://doi.org/10.1016/j.ijhcs.2022.102903

Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating generative AI in mental health: Systematic review of capabilities and limitations. JMIR Mental Health, 12, e70014. https://doi.org/10.2196/70014

Internal source foundation

The following internal Hii documents informed the case description and proposed method. They are institutional working sources, not external validation.

- The Preserved Relationship as the Unit of Analysis: AGI-26 presentation proposal, June 2026.
- HII Research Methods: Epistemic Guardrails, July 2026.
- Hii Evidence and Architecture Crosswalk v1.0, July 2026.
- Hii Longitudinal Human-AI Relationship Study: Concept and Build Requirements v0.1, July 2026.
- HII Research Reliability Standard v1.0, July 2026.