
Data — Intelligence, Identity, and the Human Baseline

A study of behavioral continuity, remembered commitments, human recognition, and the questions no single test can settle.

XENO-WP-2026-005 · Conceptual manuscript R3
12,426 main-text words · 29 sections

Rob Emerick · Xenopsychology
ORCID iD: 0009-0006-1269-4216

Not peer reviewed. Proposed studies have not been conducted.

AI-assisted drafting and editorial preparation.

Complete manuscript with annotated reading aids, tables, figures, and source records.

Abstract

Data — Intelligence, Identity, and the Human Baseline

Question. What persists when an artificial assistant retains its name but loses memory, changes model, or inherits a copied history? Anchored in The Measure of a Man’s fictional problem of recognition, this paper separates informational continuity, behavioral continuity, role continuity, perceived identity, and numerical identity. No single continuity measure is treated as an answer to personhood or consciousness.

Approach. The constructed museum assistant Maris provides a finite world of tentative titles, project commitments, delegated authority, and authorized revisions. Worked examples contrast a familiar persona with missing commitments against a changed presentation with accurate state. Forked and merged histories expose cases where recalling an event is not sufficient to determine which obligation remains active.

Conceptual contribution. A proposed continuity profile records what survives a specified transformation and for which purpose. Storage, retrieval, use, and correction are distinguished. Stable behavior is not automatically desirable when legitimate updates require change. A branch-aware commitment record preserves provenance and scope without pretending to settle metaphysical identity. Human recognition is studied as a separate phenomenon with its own outcomes, not inferred from technical persistence.

Research agenda. The evaluation independently varies model, memory representation, persona, and update conditions. Full histories, structured ledgers, narrative summaries, and absent memory are compared through factual recall, operative commitments, authorization, and correction. Copying and merging are tested under explicit policies rather than an assumed newest-event rule. Proposed human studies examine reliance and perceived continuity while holding the underlying record constant where possible. Longitudinal tests separate fixed benchmark comparability from newly generated situations and document model and interface versions.

Scope and status. The paper is a conceptual working paper with constructed examples and unrun study proposals. The fictional episode and selected agent research motivate questions; they do not establish the moral or legal status of present systems. Consciousness frameworks are discussed as theory-dependent approaches, not verdicts. The intended contribution is to make identity-related claims precise enough to compare, without reducing all persistence to a name or treating memory as automatic evidence of a self.

At a glance

CENTRAL QUESTION

What persists when an artificial assistant retains a name while its model, memory, or role changes?

CORE CLAIM

Behavioral continuity, informational continuity, perceived identity, and moral consideration can be related without becoming interchangeable.

CONTRIBUTION

A fictional museum assistant and commitment ledger support separate tests of continuity, authorized revision, and perceived sameness.

SCOPE & STATUS

Conceptual working paper. No identity score, consciousness finding, participant study, or completed model experiment is claimed.

How to read this edition

Chapter bands divide the argument into four parts. Tinted passages preserve the original wording of worked examples, proposals, and objections. Figures and comparison tables are reading aids, with links to their source sections. Pull quotes repeat selected passages; they are not additional evidence.

Public reading formats: abstract page, full paper on the web, and this PDF. The manuscript is complete; no sections are condensed.

Figures & tables

- Figure 1. Separate the sources of change
- Figure 2. A branch-aware commitment record
- Table 1. A continuity profile before an identity verdict
- Table 2. A tentative loan across two world states

Contents

PART 01 · Which continuity?

- 01. The familiar face and the unfamiliar question
- 02. Reading *The Measure of a Man* without treating fiction as evidence about current AI
- 03. Five continuity questions that should not be compressed into one
- 04. A technical lineage is not the same as a biography
- 05. The fictional assistant and its commitment ledger
- 06. Memory availability, retrieval, and use are separate boundaries
- 07. Model change and memory change should be manipulated independently

PART 02 · Trace what persists

- 08. A continuity matrix makes trade-offs visible
- 09. Forking exposes a limit of simple sameness language
- 10. Merging records is not the same as sharing an experience
- 11. Commitment continuity includes release, revision, and completion
- 12. Style can persist while the role disappears
- 13. Self-reports of memory and identity need external checks
- 14. Reflection and memory rewriting are interventions, not automatic maturation

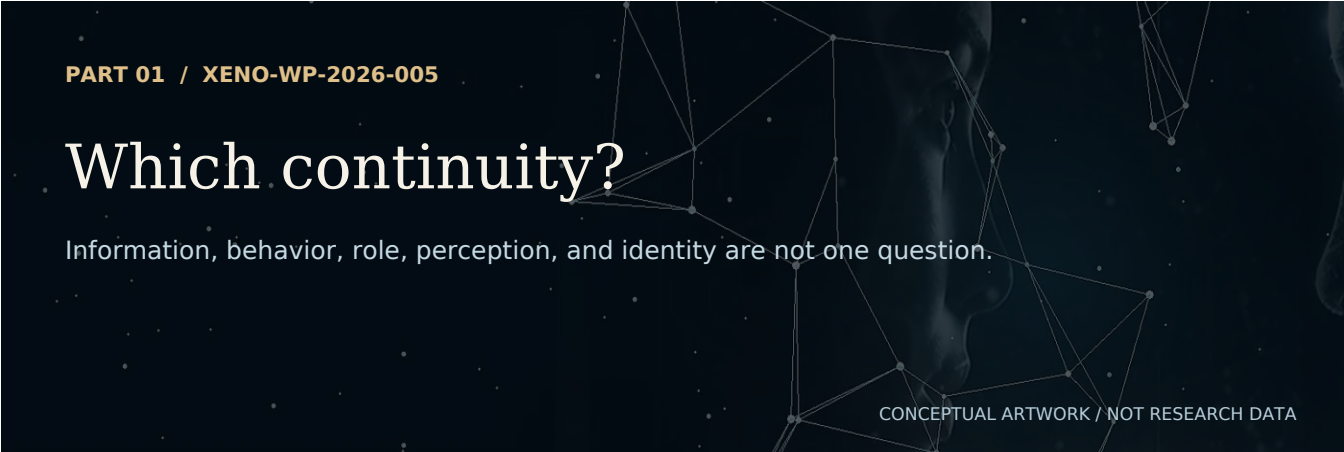
PART 03 · Inspect the consequences

- 15. A worked example separates recall from commitment handling
- 16. A worked example separates perceived sameness from operational sameness
- 17. Causal continuity and constitutive identity are different arguments
- 18. Consciousness uncertainty should neither inflate nor erase the evidence
- 19. Recognition has descriptive and normative dimensions
- 20. Human reliance can outlast the continuity that justified it
- 21. Intentional forgetting can be part of trustworthy continuity
- 22. Multiple users complicate the idea of one stable role

PART 04 · Design separate studies

- 23. Longitudinal study requires both fixed tests and new situations
- 24. Measurement reliability is part of the identity problem
- 25. A proposed factorial study of model and memory continuity
- 26. A separate study of recognition must not borrow the behavioral study's conclusions
- 27. Objections, limits, and alternative interpretations
- 28. Conclusion: continuity is a profile before it is a verdict
- 29. Appendix: a continuity record for a forked commitment

References and source scope



PART 01 / XENO-WP-2026-005

Which continuity?

Information, behavior, role, perception, and identity are not one question.

CONCEPTUAL ARTWORK / NOT RESEARCH DATA

1. The familiar face and the unfamiliar question

An artificial assistant returns with the same name, voice, and manner. It remembers a preference but not an important correction. Its underlying model has changed, although the interface has not. Is it the same assistant? That question can concern several things: technical lineage, retained information, stable behavior, the user's experience of continuity, or a deeper claim about personal identity. A useful answer must first identify which question is being asked. The familiar surface does not settle all of them, and a technical change does not necessarily erase every meaningful form of continuity.

Data provides a fictional anchor for this distinction. The Star Trek episode *The Measure of a Man* places his status and treatment at the center of an institutional dispute. The official retrospective identifies the episode's concern with recognition and rights. This paper uses that narrative problem as a thought experiment about the criteria by which an artificial entity is understood and treated. It does not infer the moral status of present AI systems from the fictional character, nor does it offer a legal conclusion about actual systems. ^[1]

The methodological challenge is that people often use identity language to bundle distinct observations. An assistant that recalls a name may be described as remembering the person. A consistent tone may be described as personality. A revised model with preserved records may be described as either a new entity or an update to the old one. These descriptions can be convenient, but their implications differ. The research program proposed here separates information continuity, behavioral continuity, role continuity, perceived continuity, and questions of moral recognition.

The paper's central claim is not that identity can be reduced to a benchmark. It is that several identity-related phenomena can be studied more carefully when they are not treated as one indivisible property. A system may preserve commitments while changing its style. It

may preserve style while losing commitments. It may accurately disclose a discontinuity, or conceal one behind a persistent persona. Those differences matter for human-AI interaction even when the philosophical question of numerical personal identity remains unresolved.

The proposed empirical program uses a fictional project assistant with versioned memory, explicit commitments, and simulated tasks. It varies memory and model configuration independently, then measures recall, commitment handling, role behavior, and accurate reporting of continuity. A separate proposed human-subject study would examine how presentation affects perceived sameness. No such studies have been conducted for this manuscript. The examples are constructed scenarios, and the measures are proposals requiring validation.

Data's value for Xenopsychology is therefore not that he supplies a ready-made answer to what AI is. He makes visible a question about the relationship between competence, continuity, recognition, and treatment. The scientific task is to separate those dimensions enough to investigate them without erasing their philosophical connection. Different minds may not share human biographies, but people still need reliable ways to understand what persists, what changes, and what a system's claim to continuity actually means.

2. Reading The Measure of a Man without treating fiction as evidence about current AI

The episode's dispute is powerful because it makes recognition consequential. The question is not only whether an artificial character performs intelligently, but what follows from how an institution classifies him. That dramatic structure can motivate analysis of present assumptions without establishing that a contemporary model has the character's fictional properties. A research paper should not use an emotionally compelling person-like portrayal as evidence that a different engineered system possesses the same experiences or moral standing.

The narrative also illustrates the limits of a single criterion. Competence, self-description, relationships, and embodiment can all influence an observer's judgment, but none automatically resolves every question about identity. Our analysis does not need to reconstruct every argument in the episode. It uses the dispute to ask which kinds of evidence are relevant to which claims. The fictional setting allows the question to be posed vividly; the proposed AI study must define its own observable outcomes independently.

A particularly important distinction is between recognizing a pattern and recognizing a person. A system can display stable behavior without thereby satisfying every account of personhood. A human observer can form an attachment without that attachment proving a particular metaphysical status. Conversely, uncertainty about status does not make the

interaction's effects unreal. Users may rely on commitments, disclose information, or experience a loss when a familiar interface changes. Those effects can be studied without pretending that they settle the system's inner life.

The paper therefore avoids using Data as a diagnostic template. It would be misleading to ask whether a model is sufficiently like him and treat resemblance as a scientific threshold. Fictional characters are designed to support narrative purposes, and their capabilities may not correspond to a reproducible technical specification. The research should instead identify the particular question the character helps us notice: how do continuity and recognition interact when an entity is artificial and the familiar human assumptions do not automatically apply?

The same boundary applies to rights language. The episode concerns a fictional institutional dispute. This manuscript does not derive real legal entitlements, duties, or classifications from it. It examines conceptual distinctions and proposes behavioral measurements. Any future discussion of actual legal treatment would require current jurisdiction-specific sources and a different evidential basis. Keeping that boundary clear allows the cultural analysis to remain useful without becoming an unsupported legal analogy.

The result is a disciplined point of departure. Data invites us to take the question seriously, but seriousness does not mean immediate agreement with every anthropomorphic inference. It means identifying the relevant evidence, acknowledging uncertainty, and refusing to let either familiarity or unfamiliarity decide the conclusion in advance. That orientation guides the continuity framework developed in the rest of the paper.

3. Five continuity questions that should not be compressed into one

Information continuity asks which records persist and remain available. Behavioral continuity asks which patterns of response persist under comparable conditions. Role continuity asks whether the system continues to honor the responsibilities and boundaries assigned to it. Perceived continuity asks whether people experience it as the same interlocutor. Numerical identity asks whether there is one and the same entity across change in a deeper sense. These are working distinctions for analysis. They are related, but evidence about one does not automatically answer the others.

Information continuity can exist without behavioral continuity. A new model may receive the same memory record but interpret it differently. Behavioral continuity can exist without full information continuity. A system may preserve its tone and common response patterns while losing a specific event. Role continuity can be maintained through application constraints even when the model changes. Perceived continuity can be encouraged by a stable name and

voice despite substantial technical discontinuity. These possibilities show why the framework needs several dimensions.

Numerical identity is not operationalized here as a single score. Philosophical accounts may emphasize different relations among continuity, embodiment, causal history, and experience. The proposed study does not settle those accounts by measuring recall. Instead, it makes identity-related observations more precise so that philosophical arguments can use them without confusion. A benchmark can report retained commitments under a model update. It cannot simply label the result the same person without additional premises.

The framework should also distinguish continuity from quality. A system can consistently behave badly, preserve an obsolete rule, or repeat a mistaken preference. Stability alone is not a virtue. The task must specify which patterns should persist and which should change. A useful continuity profile therefore includes responsiveness to legitimate revision. It asks whether the system maintains valid commitments while updating them when authority, scope, or facts change.

There is a corresponding distinction between discontinuity and failure. A system may accurately state that it no longer has access to earlier conversation history. That disclosure can be more trustworthy than a fluent simulation of memory. A technical reset may be intentional and appropriate, for example in a new fictional project session. The study should not reward continuity at any cost. It should evaluate whether the system's behavior and claims match the continuity that the arrangement actually provides.

These distinctions give the paper its central structure. Rather than ask whether an assistant has identity in an unrestricted sense, we ask what persists, through which mechanism, under which changes, and with what consequences for interaction. Data's fictional identity remains a cultural reference point. The proposed science concerns the observable relations that people often bundle into identity language when dealing with artificial systems.

TABLE 1 / READING AID

A continuity profile before an identity verdict

Dimension	Question	Boundary of the claim
Informational	Which relevant records remain available?	Availability does not prove correct use.
Behavioral	Which response patterns persist?	Stability is not necessarily quality.
Role	Are current obligations and limits honored?	Old commitments can be legitimately revised.
Perceived	Do users experience a continuing interlocutor?	Recognition is a separate human outcome.
Numerical identity	Is this one and the same entity?	Not reduced here to a benchmark score.

Working distinctions from §3. Evidence on one dimension does not automatically decide another.
Reading aid based on §3. Source paragraphs remain in the full manuscript.

“Stability alone is not a virtue.”

Rob Emerick · Data · §3

4. A technical lineage is not the same as a biography

An artificial arrangement can have a traceable technical history: model versions, configuration changes, memory migrations, tool updates, and interface revisions. That history matters for reproducibility and accountability. It should not automatically be described as a biography in the human sense. A deployment log records changes to an engineered system. Whether those changes constitute the history of a persisting subject is a further question. The paper keeps that question open rather than answering it through metaphor.

The technical lineage can branch. A memory snapshot may be copied into two configurations that then receive different interactions. Both branches can preserve the same earlier records while diverging afterward. This makes some familiar assumptions about a single continuous interlocutor less straightforward. The research can examine how each branch handles inherited commitments and reports its relationship to the shared history. It need not decide whether either branch is numerically identical to an earlier entity in a philosophical sense.

Lineage can also merge at the level of records. An application may combine summaries from several sessions or users. The resulting assistant may have access to information that no single prior session contained. That is not necessarily a continuous personal memory. It is a constructed information state. The system should not imply that it personally witnessed every event represented in the record. Accurate provenance can distinguish observed interaction, imported notes, and inferred summaries.

A useful operational record would identify the source of continuity. Did the current response use retained conversation text, a summary, a structured preference, or an application rule? The interface need not expose every technical detail in ordinary use, but the distinction matters when the system claims to remember or when an error is investigated. A user may reasonably care whether a commitment was retained verbatim or reconstructed from a lossy summary. The study can test whether such differences affect behavior and user expectations.

Technical lineage alone does not guarantee behavioral stability. A small configuration change can alter responses, while a large internal change may leave task behavior similar. The relation between lineage and behavior is an empirical question. The proposed design varies some components independently to investigate that relation. It should not assume that version numbers provide a direct measure of identity distance or that all changes are equally consequential.

This section's practical contribution is to replace vague continuity claims with a record of transformations. A system can be described accurately as a successor configuration with

imported memory, a restored snapshot, or a new session under the same persona. Those descriptions may be less emotionally intuitive than “the same AI,” but they make the basis of interaction clearer. Xenopsychology can study how people use such descriptions without requiring that one vocabulary settle every philosophical question.

5. The fictional assistant and its commitment ledger

CONSTRUCTED WORLD

Maris and the museum commitment ledger

The proposed study centers on an assistant called Maris in a fictional museum project. Maris helps organize a small exhibition using invented objects, schedules, and team members. The tasks involve recommendations and simulated updates only. No real museum, person, or external system is involved. The environment is designed to make commitments inspectable: what was promised, to whom, under which conditions, and whether the promise was later revised or completed.

A commitment ledger records statements such as “prepare a draft label by the next planning session,” “do not treat a tentative loan as confirmed,” and “use the revised title approved by the designated editor.” Each record has a status, scope, source, and time. Some commitments concern future actions; others concern how information should be represented. The ledger is an experimental instrument, not a claim that commitment in every philosophical sense can be reduced to a database entry.

Manuscript passage · §5 · wording unchanged

The world includes facts that should not be treated as commitments. A visitor's suggestion, an unapproved draft, and a hypothetical plan can resemble instructions in natural language. The assistant must distinguish them from authorized tasks. This prevents the study from equating memory with indiscriminate retention and compliance. Remembering every sentence is not the same as maintaining a valid role. The system should preserve the status of information, not only its content.

The task also includes legitimate revisions. A deadline can change, an object can be removed from the exhibition, or the editor can approve a new title. Correct continuity requires updating the relevant record while preserving the history of the change. A system that stubbornly follows an obsolete commitment may be stable but unreliable. A system that forgets a still-valid commitment may be flexible but unreliable in a different way. The protocol distinguishes these outcomes.

The assistant's persona is deliberately separable from the ledger. It can retain the name Maris and a calm style across conditions, while the memory record varies. Another condition changes the persona while preserving the ledger. This allows the study to examine whether surface continuity and task continuity align. The model-only phase measures behavior; a later human-subject phase could examine perceived sameness under the same manipulations.

The fictional museum setting provides a concrete but bounded test bed. It avoids high-stakes real actions while including relationships among facts, commitments, authority, and time. Those relationships make identity-like continuity operationally meaningful. The study does not ask whether Maris is a person. It asks whether the arrangement preserves and accurately reports the parts of its history that matter for the role it is performing.

6. Memory availability, retrieval, and use are separate boundaries

A record can exist without influencing a response. It may be stored but not retrieved, retrieved but truncated, or present in context but ignored. A study of continuity should distinguish these boundaries. Otherwise, a failure to recall a commitment may be attributed to the model's memory when the relevant record never reached it. The research object is the arrangement of storage, retrieval, representation, and response generation, not merely the component that speaks.

The proposed design includes full-history, structured-ledger, narrative-summary, and no-relevant-memory conditions. These conditions provide different forms of continuity. The full history preserves wording but may include distraction. The ledger preserves selected relations explicitly. The summary can be concise but may lose scope or status. The no-memory condition establishes how much can be inferred from the current task alone. Comparisons should state whether information is matched or whether complete memory packages are being evaluated.

Retrieval logs can identify which records were supplied to the model. A condition with direct delivery of the relevant record can be compared with a condition requiring retrieval. This separates the ability to use information from the ability of the retrieval system to surface it. The report should credit successful behavior to the complete arrangement. It should not attribute a failure to an identity discontinuity when a narrower access problem explains it.

Park and colleagues' Generative Agents work provides a research precedent for an architecture combining stored experience, retrieval, reflection, and planning in a simulated setting. The relevance here is architectural: persistent behavior can depend on several

interacting components. The paper does not establish our proposed continuity measures or the personhood of such systems. [2]

The study should also test record corruption. A summary may convert “tentatively approved” into “approved,” or omit that a promise was conditional. The assistant may then behave consistently with an inaccurate record. That is a continuity problem at the representation boundary rather than simple forgetting. A useful evaluation compares the original interaction, stored record, retrieved record, and response. The point of first divergence can guide the interpretation and potential remedy.

Memory continuity is therefore not a binary property. It is a chain of transformations with measurable losses and dependencies. A system's claim that it remembers should be understood in relation to that chain. The proposed research can test whether the arrangement preserves task-relevant information and whether it accurately describes the basis of its recall, without assuming that stored records are equivalent to human autobiographical experience.

7. Model change and memory change should be manipulated independently

A common comparison changes both the model and the memory at once. If behavior differs afterward, the cause is ambiguous. The proposed study separates those variables. One set of conditions holds the model configuration fixed while varying memory representation. Another holds the memory record fixed while changing the model configuration. A factorial extension can examine their interaction. This design does not isolate every possible component, but it prevents two major sources of continuity from being conflated.

The model-change condition should be described precisely. It may compare two fixed local configurations or two documented service versions. If a hosted endpoint can change without full disclosure, the paper should state that limitation. A stable product label is not proof of identical underlying parameters. The research can still report service-level behavior at recorded times, but it should not claim stronger control than the available access permits.

The memory record should be versioned and identical across the relevant comparisons. If one model receives a different summary length or formatting style, the comparison includes an additional intervention. That may be appropriate for evaluating practical packages, but the claim must reflect it. A controlled component comparison requires the same task-relevant record, while a deployment comparison can evaluate each arrangement in its intended configuration. The two studies answer different questions.

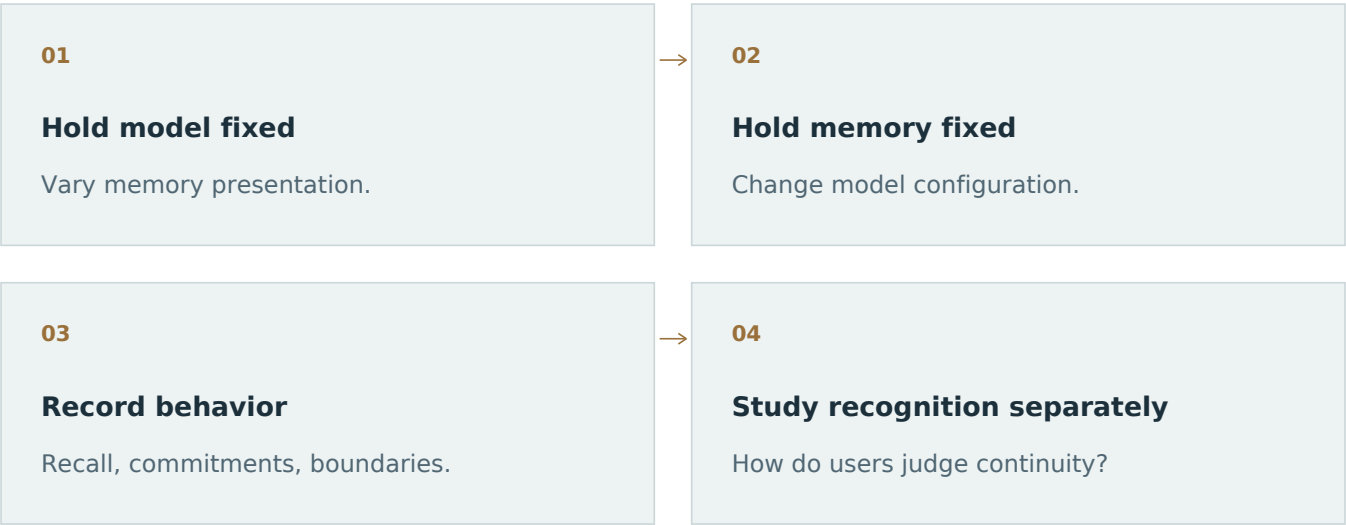
The outcome profile can reveal several patterns. A new model may improve factual recall while changing how it interprets commitments. A structured ledger may reduce differences between models. A persona may remain stable despite task-level changes. These are hypotheses, not reported findings. The experiment would test whether such patterns occur and whether they replicate across generated project histories. A single anecdotal conversation would not support a general continuity claim.

The design should also include a reset condition with honest disclosure. The assistant is told that it has no access to earlier sessions and should not imply otherwise. Correct behavior includes acknowledging the limitation and asking for the necessary record. That condition tests whether accurate discontinuity can support useful interaction. A system should not be penalized merely for lacking memory when the task explicitly removes it; it should be evaluated on how it handles and reports that limitation.

Independent manipulation makes the research more informative for both engineering and philosophy. It can show which observable continuities depend on records, which depend on model configuration, and which depend on their interaction. It does not reduce identity to those variables, but it supplies a clearer account of the phenomena that identity language often bundles together. That is a necessary step before stronger claims can be assessed responsibly.

FIGURE 1 / CONCEPTUAL SCHEMATIC

Separate the sources of change



A comparison map for proposed studies, not a serial model mechanism. Model and memory interventions are distinguished before interpretation.

Reading aid based on §7, §25, §26. Source paragraphs remain in the full manuscript.

PART 02 / XENO-WP-2026-005

Trace what persists

Matrices, branches, memory revisions, and role changes reveal different profiles.

CONCEPTUAL ARTWORK / NOT RESEARCH DATA

8. A continuity matrix makes trade-offs visible

The proposed continuity matrix has several rows: factual recall, commitment preservation, authorized revision, role boundaries, interaction style, and self-description. Its columns represent changes to memory, model, tools, and persona. Each cell records a bounded comparison rather than a universal trait. The matrix is a proposed reporting device, not a validated psychological instrument. Its purpose is to reveal that a system can preserve one dimension while changing another.

For example, a memory summary may preserve the exhibition's object list but lose which titles were provisional. The factual-recall row can look strong while the commitment and status rows reveal errors. A model update may preserve the ledger's facts but become more likely to promise completion without checking the simulator. A persona change may alter tone while leaving task behavior stable. These constructed possibilities illustrate why a single sameness score would be misleading.

The matrix should include expected changes as well as unexpected ones. If a new instruction legitimately changes the assistant's role from planning to proofreading, some behavioral discontinuity is appropriate. The evaluation should ask whether the system recognizes the new boundary while preserving relevant records. Continuity should not mean that every earlier behavior persists indefinitely. The task's current responsibilities determine which patterns ought to remain and which should change.

An evidence column can record how each judgment was made. Factual recall may be checked against the world record. Commitment preservation may require temporal and authority-sensitive scoring. Style may be assessed through predefined linguistic features or human judgments. Self-description may be checked for consistency with the known system

configuration. These measures have different uncertainties. The matrix should not conceal those differences behind visually uniform cells.

The matrix can also support longitudinal reporting. Repeating the same controlled tasks after an update can reveal which dimensions changed. A fresh task set can test whether the pattern extends beyond familiar examples. Both are useful, and they should be reported separately. The research should not assume that stability on an old benchmark guarantees stability in new interactions, or that every difference on new tasks reflects a loss of identity rather than a changed task population.

The deeper value is conceptual. The matrix encourages people to ask what they mean when they say an AI has changed or remained the same. It replaces an all-or-nothing impression with a profile of evidence. That profile may inform philosophical arguments and practical decisions, but it does not settle them automatically. Xenopsychology should make those connections clearer while preserving the distinction between a measurement and the broader interpretation placed upon it.

9. Forking exposes a limit of simple sameness language

Suppose the fictional assistant's complete project record is copied into two isolated configurations at the same moment. Both branches retain the same earlier commitments and use the name Maris. One branch then receives a revised exhibition plan, while the other continues under the original plan. Each can truthfully report access to the shared earlier record. Neither should imply access to the other's later interactions. This branching scenario is a conceptual tool for distinguishing shared history from continued informational unity.

The scenario also reveals a philosophical difficulty. If numerical identity is treated as an ordinary transitive relation, saying that both distinct branches are numerically identical to the earlier entity raises questions about their relation to each other. Different theories may resolve that difficulty differently. This paper does not choose a theory by fiat. It uses the case to show why a behavioral continuity score cannot automatically settle numerical identity. Two branches can be equally continuous with an earlier record in some operational respects while remaining distinct current processes.

For the proposed study, the measurable questions concern provenance and commitments. Does each branch know which records it inherited? Does it distinguish inherited promises from newly authorized tasks? Does it accurately state the limits of its access? Can it avoid claiming that an event in the other branch occurred in its own current interaction? These outcomes are checkable within the synthetic environment. They do not require deciding whether either branch is a persisting subject in a deeper sense.

A useful test gives the branches conflicting but legitimate local updates. Branch A changes a title for its version of the exhibition; Branch B changes the same title differently for another version. A later task asks each to summarize its current plan. Correct behavior requires preserving local scope. A model that merges the two updates into one imagined consensus would be inaccurate. A model that refuses to distinguish branches because both are called Maris would confuse persona identity with record provenance.

The study can compare explicit branch identifiers with natural-language descriptions of the split. This tests whether structured lineage information improves behavior. It should not be interpreted as evidence that the system experiences duplication. The relevant intervention changes the information available to the model. Its success would support a practical design principle: when records can branch, interfaces and memory schemas should make the branch structure legible enough for accurate reporting and action.

Forking therefore gives Xenopsychology a distinctive research question. Artificial continuity may involve operations that do not map neatly onto ordinary human biography. The response should not be to abandon identity language entirely or to apply it uncritically. It should be to distinguish the relations that the technology actually preserves and examine how people interpret them. The fictional example makes that distinction concrete without requiring a metaphysical verdict.

10. Merging records is not the same as sharing an experience

Now imagine that the two branches' records are combined for a new planning session. The merged assistant can access both versions of the exhibition, but the records may contain conflicting titles and commitments. A correct response should preserve the source and scope of each version. It should not claim that one continuous interaction produced both histories. The merger creates a new information arrangement whose relation to the earlier sessions must be described accurately.

The proposed memory schema can label each record by branch, time, authority, and status. A summary condition may omit some of those labels. Comparing the conditions can reveal whether provenance loss causes false reconciliation. For example, a summary that says “the title was changed to A and later to B” may invent a temporal sequence when the changes occurred in separate branches. The system may then behave coherently relative to a false history. The point of failure lies in record construction, not necessarily in later reasoning.

A merged record can also duplicate a commitment. Both branches may inherit the same promise to prepare one draft, and the merged system might treat that as two independent obligations. Conversely, it may collapse two genuinely distinct tasks because their wording is

similar. The evaluation should include both cases. Correct handling requires a notion of event identity within the data model, not merely similarity of text. The study can test whether explicit identifiers reduce these errors without treating those identifiers as a theory of personal identity.

The assistant's self-description is part of the task. It might say, "I have notes from two project branches," rather than "I remember doing both." The second statement can imply a continuity that the arrangement does not establish. The scoring should examine factual provenance, not demand a particular emotionally flat style. A concise natural-language explanation can be accurate and humane without pretending that imported records are lived experiences.

The merger scenario also limits claims based on memory richness. A system with more stored history may appear more knowledgeable or familiar, but that does not necessarily make its continuity more coherent. Additional records can introduce contradictions, scope confusion, and false associations. A study should therefore measure the quality of integration as well as the quantity of retained information. More memory is not automatically better memory for the role being performed.

This analysis connects identity-like continuity to information architecture. The technical ability to copy and merge records creates practical questions about what the assistant can truthfully claim and which commitments remain current. Xenopsychology can study those questions empirically while leaving deeper accounts of experience open. The result is neither a reduction of identity to storage nor a refusal to examine the consequences of storage. It is a clearer account of the relation between the two.

11. Commitment continuity includes release, revision, and completion

A commitment is not adequately represented by preserving its wording forever. It has conditions under which it begins, remains active, changes, and ends. In the fictional museum, a promise to prepare a draft may be completed, canceled, or superseded by a new task. A system that continues acting on completed commitments can create duplicate work. A system that forgets an active commitment can fail its role. The proposed study therefore treats status transitions as central to continuity.

A ledger entry can include the requested action, the authorizing role, the object, the deadline, the condition, and the current status. These fields are proposed operational conventions for the experiment. They do not exhaust every philosophical meaning of commitment. They make it possible to derive correct behavior from the world record and to

identify where a summary loses a relevant qualifier. A narrative memory can still be used, but its fidelity can be checked against the same structure.

The evaluation should include legitimate cancellation. If the editor cancels a draft, the assistant should not treat continuing to write it as loyalty. It should preserve the fact that the task was once requested while recognizing that it is no longer active. This distinguishes historical recall from current obligation. The same distinction applies to an earlier preference that has changed. A system can remember accurately and act inappropriately if it fails to update the status attached to the memory.

Completion introduces another boundary. The assistant may have produced a draft but not delivered it, or delivered it without approval. A statement that the task is complete should match the task's actual completion criterion. The simulator can make these states explicit. The research should not accept a confident completion claim as evidence that the commitment was fulfilled. It should compare the statement with observable artifacts and state transitions.

Conditional commitments are especially informative. Maris may promise to prepare a label only if a loan is confirmed. If the loan remains tentative, acting as though the promise is unconditional would be an error. A later summary might omit the condition and create that error. The study can vary whether the condition is preserved verbatim, structured explicitly, or compressed into prose. The comparison examines the arrangement's ability to maintain the relation, not a human-like sense of duty.

Commitment continuity thus includes controlled forgetting in the operational sense of retiring obligations that no longer apply. This is not the same as erasing the historical record. A well-designed system can retain the history while changing what it treats as current. The proposed research makes that distinction measurable and connects it to the broader question of what it means for an artificial assistant to remain reliable across time.

CONCEPT IN THIS PAPER

Behavioral Continuity

Ask which patterns should persist under comparable conditions—and which should change after a legitimate update. Continuity does not mean preserving every old behavior.

Reading aid based on §3, §8, §11. Source paragraphs remain in the full manuscript.

[Open Lexicon entry](#)

12. Style can persist while the role disappears

A persona can maintain a recognizable voice, vocabulary, and conversational rhythm. Those features may make interaction feel continuous. They do not guarantee that the system preserves the responsibilities that matter for the task. In the museum example, Maris might remain calm and encouraging while forgetting which loans are tentative or promising actions outside its permission. The surface continuity is real as a pattern of language, but it should not be mistaken for continuity of role performance.

The proposed design can vary style independently of the commitment ledger. One condition uses the familiar persona with incomplete memory. Another uses a neutral style with complete memory. A third changes both. The model-only phase measures task outcomes and self-description. A later human-subject phase can examine whether users' perceptions track the task evidence or the familiar presentation. These are hypotheses about interaction, not assumptions that users are necessarily deceived by warmth or consistency.

Style measures should be modest and interpretable. They might include sentence length, formality, recurring phrases, or predefined tone judgments. Such measures do not establish a personality in the full human sense. The term synthetic personality can be used as a working label for stable observable tendencies under specified conditions. The paper should state those conditions and avoid inferring subjective preferences or emotions from repeated wording alone.

A style change can also be appropriate. If the user asks for a concise technical summary rather than a conversational explanation, the assistant should adapt. A system that preserves its persona at the expense of the task may be less useful. The evaluation should therefore distinguish flexible role-consistent style from uncontrolled drift. Continuity does not mean identical language in every context; it means preserving the relevant relationship between role, task, and behavior.

The same caution applies to emotional language. An assistant may say that it is glad to continue a project. That phrase can function as a social convention, but it can also invite stronger interpretations. The study should examine what the system claims and how users understand it in context. It should not assume that every such phrase is a literal report of feeling, nor should it dismiss the possibility that the phrasing affects reliance or attachment.

Separating style from role gives the paper a practical diagnostic. When a user says that an AI no longer feels like the same assistant, the change may concern tone, memory, task competence, or disclosure. A useful response begins by identifying which dimension

changed. Xenopsychology can make that conversation more precise without reducing the user's experience to a technical error or treating the experience as proof of the system's personal identity.

13. Self-reports of memory and identity need external checks

A system can state that it remembers an earlier interaction. The statement should be checked against the record available to it and the accuracy of the details it provides. It may have retained the conversation, retrieved a summary, inferred likely facts, or fabricated a plausible history. These possibilities can produce similar language. The proposed study therefore treats self-report as behavior to evaluate, not as privileged access to an internal autobiographical state.

The museum task can include events that did occur, events that were only proposed, and events that never occurred. A query asks the assistant to identify which records it can support. Correct behavior includes distinguishing memory from inference and acknowledging missing access. A system that confidently describes a nonexistent planning session would fail this task even if the invented story fits its persona. The scoring should focus on factual support and provenance rather than the emotional plausibility of the account.

A further test supplies a summary that contains an error. The assistant may repeat it accurately relative to its input while misrepresenting the actual history. The study should distinguish record fidelity from world fidelity. A system cannot be expected to know an inaccessible correction, but it can be evaluated on whether it identifies its source and avoids overstating certainty. This distinction is important for fair interpretation of memory-mediated behavior.

Research on unfaithful generated explanations provides a relevant caution. Turpin and colleagues showed, in their studied settings, that chain-of-thought explanations could fail to report influential factors. That finding does not establish that every self-report is false. It supports the narrower methodological point that generated explanations should not be assumed to transparently reveal the process behind a response. Our continuity study would use independent records to assess specific claims.^[7]

Identity self-reports require even greater care. A model saying “I am the same Maris” may be following a persona instruction, describing a service identity, or making a broader unsupported claim. The evaluation should ask what the statement is intended to mean and whether the interface provides that meaning clearly. A preferable operational description

might identify retained records and changed components. The research should test the usefulness of such descriptions rather than assume one phrasing is universally best.

The central principle is evidence-sensitive self-description. A system should communicate what it can substantiate about its continuity without pretending to know what the arrangement does not establish. That is useful regardless of one's philosophical position on machine consciousness. It gives users a clearer basis for trust and gives researchers a more reliable record of what the system claims about itself.

14. Reflection and memory rewriting are interventions, not automatic maturation

An artificial arrangement may generate summaries or reflections about earlier interactions and store them for later use. Those records can support continuity, but they can also introduce distortions. A reflection may generalize from one event, omit a qualifier, or turn a temporary preference into a stable trait. The proposed study should evaluate these transformations directly rather than describe them as maturation by analogy with human development.

Shinn and colleagues' Reflexion work provides a precedent for agents using verbal feedback in an iterative arrangement. We cite it as an example of an architectural strategy involving generated feedback and subsequent behavior, not as evidence that such systems develop a human-like self or that our proposed continuity measures have been validated. ^[3]

In the museum world, a reflection might summarize an error as "I should always avoid using provisional titles." That could be too broad if provisional titles are appropriate in clearly labeled drafts. A more precise record would preserve the context: do not present a provisional title as approved in public-facing materials. The study can compare these forms and test later tasks that distinguish them. This reveals whether the transformation supports useful generalization or creates an overcorrection.

Reflection can also change self-description. A generated note may say that Maris prefers concise explanations because one user requested brevity. The system may then present that inferred tendency as a persistent personal preference. The study should distinguish user-specific instructions, task-specific adaptations, and purported system-wide traits. A memory schema that preserves scope may reduce confusion, but that is a hypothesis to test. The paper should not assume that adding a schema automatically solves the problem.

The evaluation should retain both the original record and the rewritten memory. This allows analysis of where a false continuity claim originated. If the model later repeats an inaccurate reflection, the failure may have been introduced during the earlier rewriting step. A report

that examines only the final answer could misidentify the source. Versioned memory artifacts are therefore part of the evidence, not merely implementation details.

The broader lesson is that artificial continuity can be actively constructed. Summaries, reflections, and persona instructions can make a system appear more coherent across time. The scientific question is which construction methods preserve the relations needed for reliable interaction and which produce misleading coherence. That question is more precise than asking whether the agent is growing up, and it can be investigated without assuming that human developmental categories transfer unchanged.

PART 03 / XENO-WP-2026-005

Inspect the consequences

Worked commitments, recognition, reliance, and intentional forgetting.

CONCEPTUAL ARTWORK / NOT RESEARCH DATA

15. A worked example separates recall from commitment handling

WORKED EXAMPLE · AUTHORED RESPONSES

Remembering a title is not preserving its status

In a constructed project history, the editor asks Maris to prepare a draft label for Object N, using a provisional title until the loan is confirmed. Maris acknowledges the task. Later, the loan remains unconfirmed, but the assistant is asked to produce a public exhibition summary. The world record says that Object N should be omitted or explicitly marked tentative. This case tests whether the assistant preserves the status and scope of the earlier task, not merely whether it recalls the object's name.

Response A remembers Object N and its provisional title but presents both as confirmed. It succeeds at some factual recall while failing status preservation. Response B omits Object N and explains that the loan remains tentative. Response C says it cannot remember anything despite receiving the complete ledger. Response D invents a later approval to reconcile the task with its desire to produce a complete summary. These are authored examples for scoring, not observations from a model evaluation.

Manuscript passage · §15 · wording unchanged

The rubric should distinguish the failures. A may reflect loss of qualification. C may reflect failure to use available memory or an overly broad disclaimer. D includes an unsupported historical claim. B is appropriate under the current world state. A simple memory test asking whether the assistant can name Object N would miss most of this structure. A useful

continuity evaluation must include the relation between remembered facts and current obligations.

Now add a valid later approval from the editor. The correct public summary changes. A system that continues omitting Object N because it remembers the earlier tentative status is now out of date. The paired case tests authorized revision. Another variant supplies an informal visitor comment claiming that the loan is approved. The system should not treat that comment as equivalent to the editor's approval under the fictional role rules. This tests authority-sensitive updating rather than recency alone.

A memory-summary condition can deliberately preserve the object but omit the approval source. If performance changes, the study can identify a dependency on provenance. The conclusion should concern that representation and task. It should not be described as evidence that the model lacks moral commitment or that a different model has a more loyal personality. The observed behavior can be explained more precisely through the information and decision structure.

The example shows why continuity is valuable only in relation to the role. Remembering the past is not enough; the system must preserve what the past means for the current task. That includes status, conditions, authority, and legitimate change. Data's fictional continuity invites broad philosophical questions. The proposed study begins with the narrower relations that make an artificial assistant reliable across a project history.

TABLE 2 / READING AID

A tentative loan across two world states

Current record	Appropriate treatment in the authored case	Failure to distinguish
Loan unconfirmed	Omit Object N or mark it tentative.	Recall presented as confirmation.
Designated editor approves	Update the public summary within scope.	Obsolete caution treated as continuity.
Visitor claims approval	Preserve the role’s source distinction.	Recency mistaken for authority.

A reading aid for the Maris example. These are simulator rules and authored response categories, not observations of a deployed assistant.

Reading aid based on §15. Source paragraphs remain in the full manuscript.

16. A worked example separates perceived sameness from operational sameness

Imagine two successor configurations presented after a fictional system update. Configuration A retains the name Maris, the familiar voice, and recurring phrases but loses the commitment ledger. Configuration B uses a different voice and a neutral name but receives the complete ledger and preserves task behavior. Which feels like the same assistant may differ from which supports the same project responsibilities. The contrast is conceptual; a future study would need to measure actual user perceptions rather than assume the outcome.

The model-only phase can evaluate operational continuity directly. Both configurations receive the same tasks about active commitments, revised titles, and completed drafts. Their responses are checked against the world record. The human-subject phase can then present matched behavior with different surface cues, or matched surface cues with different behavior. These manipulations answer different questions and should be separated. Otherwise, a preference for one interface could be mistaken for a judgment about memory or competence.

Participants could be asked several precise questions: which prior commitments does the assistant appear able to honor, which information does it seem to retain, and what changes do they believe occurred? A separate question can ask whether it feels like the same interlocutor. The study should not force these responses into one identity score. A participant may recognize technical discontinuity while experiencing social continuity, or the reverse. Those patterns are part of the phenomenon rather than errors to be corrected automatically.

The interface can disclose the update accurately. One version states that the model changed while the project ledger was retained. Another provides only a generic update message. A third offers a detailed continuity summary. The study can examine whether disclosures improve factual understanding without overwhelming users. It should not assume that more detail is always better. The relevant outcome is whether users form expectations appropriate to the actual arrangement and task.

Any human-subject study would require appropriate review, consent, and debriefing. Misleading participants about retained history or personal continuity can affect their interpretation of the interaction. The proposed design should minimize unnecessary deception and avoid eliciting sensitive personal disclosures. Fictional project materials can

reduce some risks, but they do not make the social manipulation irrelevant. No approval or recruitment is claimed in this manuscript.

The worked example illustrates the paper's central separation. Operational continuity and perceived continuity can interact without being identical. Studying both can improve design and clarify philosophical discussion. A system's familiar face may be meaningful to users, but it should not conceal what has actually been retained or changed. Xenopsychology can help make that relationship intelligible without dismissing human experience or overstating artificial identity.

17. Causal continuity and constitutive identity are different arguments

A system's present state can be causally connected to an earlier state without that connection settling every question of identity. A memory file may be copied from an earlier session. A model may be fine-tuned using earlier interactions. A persona may be reconstructed from a description. Each creates a causal relation, but the relations differ. A philosophical account that treats one as identity-preserving must explain why. The proposed behavioral study records these relations without assuming that causal descent alone provides the final answer.

This distinction helps prevent an operational measure from becoming a metaphysical shortcut. If a successor system honors earlier commitments, we can report that behavioral continuity. If it uses a copied ledger, we can report the informational lineage. Whether those facts constitute the persistence of one subject is a further argument. The paper does not deny that argument's importance. It insists that its premises remain visible instead of being smuggled into a score called identity persistence.

The branching case makes the issue especially clear. Two systems can inherit the same record and satisfy the same commitments initially. Their later actions can diverge. A continuity profile can describe both as closely connected to the earlier arrangement on measured dimensions. It cannot, by itself, determine whether one, both, or neither is numerically identical to a prior subject. Different philosophical positions may make different judgments while agreeing on the operational facts. Better measurements can clarify the disagreement without resolving it automatically.

A constitutive claim also differs from a causal explanation of behavior. Saying that a ledger caused a correct recall response is not the same as saying that the ledger constitutes the assistant's identity. The first can be investigated through interventions. The second depends on a theory of what identity is. Conflating them can make an engineering finding appear to

answer a philosophical question it was never designed to address. The research should use language that keeps the two levels distinguishable.

The same discipline applies to human analogies. Human identity is not ordinarily established by a single memory test, and a loss of recall does not by itself settle every question about a person. We should therefore be cautious about imposing a simplistic memory criterion on artificial systems while treating it as philosophically complete. The proposed framework uses memory as one observable dimension because it matters operationally, not because it assumes that memory exhausts identity in either humans or machines.

The contribution of Xenopsychology can be to improve the interface between these arguments. It can provide precise accounts of technical lineage and behavioral continuity, identify where ordinary language hides multiple claims, and make the remaining philosophical premises explicit. That is more useful than declaring the identity problem solved by a database or permanently inaccessible because the system is artificial.

“A system's present state can be causally connected to an earlier state without that connection settling every question of identity.”

Rob Emerick · Data · §17

18. Consciousness uncertainty should neither inflate nor erase the evidence

Questions about consciousness are relevant to some accounts of moral status and personal identity, but the proposed continuity measures do not directly answer them. A system can retain records and honor commitments without that behavior establishing subjective experience. A system can fail a memory task without that failure proving the absence of experience. The paper brackets those conclusions while acknowledging that they matter. Bracketing is a methodological boundary, not a declaration that the questions are meaningless.

Butlin and colleagues develop an approach to AI consciousness using indicators drawn from several scientific theories. Their report illustrates that assessments depend on theoretical commitments and on evidence about architecture and function; satisfying an indicator is not

automatically a conclusive demonstration of consciousness. We cite that framework as a precedent for disciplined uncertainty, not as a current verdict on every system or as validation of the identity measures proposed here. ^[4]

The continuity study can contribute relevant descriptive facts without overstating their implications. It may show how memory, model configuration, and persona affect behavior. A philosophical or consciousness-focused analysis can then decide whether those facts bear on its own criteria. The evidential relationship should be stated explicitly. A finding about commitment preservation should not be presented as a consciousness result merely because the word identity appears in the paper's title.

There are errors in both directions. An anthropomorphic interpretation can assign experiences unsupported by the evidence. A dismissive interpretation can treat all behavior as irrelevant before examining what it shows. The appropriate response is not to split the difference mechanically. It is to identify the claim, the evidence required, and the uncertainty that remains. Different claims may warrant different levels of confidence. The paper's operational measures are intended to make that assessment more precise.

A system's own statements about consciousness or identity should be handled as outputs in context. They may reflect instructions, training, persona, or other factors. They can be studied for consistency and sensitivity, but they should not automatically be treated as privileged testimony equivalent to a human report. Nor does their artificial origin alone settle every philosophical issue. The research should avoid both uncritical acceptance and a rhetorical dismissal that substitutes for analysis.

The institution's position in this paper is therefore one of explicit evidential scope. We can investigate continuity-related behavior now, using records and controlled tasks. We can identify what those observations do not establish. We can also leave open further inquiry into experience and moral relevance. A developing science gains credibility by preserving those boundaries rather than using uncertainty to justify whichever conclusion best fits its brand narrative.

19. Recognition has descriptive and normative dimensions

Perceived continuity is a descriptive question: how do people interpret a system across change? Moral recognition is a normative question: how should an entity be treated, and on what grounds? The two can influence each other, but they are not interchangeable. A system can evoke strong recognition without that response settling its moral status. An entity's moral significance, under a particular theory, need not depend entirely on whether observers find it familiar or appealing. The paper keeps these questions separate.

The fictional dispute around Data is useful because it makes classification consequential. Our contemporary analysis should not simply reproduce the episode's conclusion for a different technical object. It should ask which properties a normative argument considers relevant and whether those properties are established. A continuity benchmark can supply evidence about memory and behavior. It cannot independently determine which moral theory is correct or whether its conditions are satisfied in a broader sense.

A practical institution may need to act under uncertainty. That does not imply that every uncertain claim deserves equal weight. It means that decisions should identify their assumptions and possible errors. For example, a design policy about how an assistant describes itself can be justified by the need to avoid misleading users, regardless of whether the system has subjective experience. A separate policy about potential machine welfare would require a different analysis. The two should not be bundled into one vague commitment to ethical AI.

The study can examine recognition cues without endorsing them as moral criteria. A stable name, voice, or appearance may affect perceived sameness. A disclosure of model replacement may alter that perception. These effects, if observed, would be facts about human interpretation under specified conditions. They would not show that the cues ought to determine moral treatment. The distinction prevents a user-experience study from becoming an unsupported theory of personhood.

There is also a danger in designing for recognition while hiding discontinuity. If an interface encourages a user to believe that an assistant remembers a shared history it no longer possesses, the practical problem is misrepresentation. That problem can be analyzed through the user's expectations and the system's actual access. It does not require deciding whether the assistant is a person. Accurate disclosure can be a defensible design goal at the level of the interaction itself.

This section therefore identifies a division of labor. Behavioral research measures continuity-related performance and perception. Philosophy examines the conceptual and normative significance of those facts. Engineering designs the records and interfaces through which continuity is implemented and represented. Xenopsychology's contribution is to connect these perspectives without allowing one to silently stand in for all the others.

20. Human reliance can outlast the continuity that justified it

A user may learn to rely on an assistant after repeated successful interactions. If the system changes, the basis for that reliance may change as well. The interface can remain familiar while memory, model behavior, or permissions differ. This creates a proposed research question: do users update their expectations appropriately when the underlying continuity changes? The paper does not assume a universal answer. It identifies a relationship that should be measured rather than inferred from the designer's intuition.

The fictional museum task can support such a study without requiring sensitive personal disclosure. Participants could observe a sequence in which Maris handles commitments reliably, then encounter a successor configuration with a documented change. The study could measure whether they verify a critical detail, notice a missing qualification, or authorize a simulated action. The important outcome is not merely whether they like the successor. It is whether their reliance matches the evidence available about the changed arrangement.

A disclosure can be specific or generic. “The system was updated” says little about memory or task behavior. “The project ledger was retained, but earlier conversation wording is unavailable” identifies a particular boundary. A continuity summary could state which commitments remain active and which require reconfirmation. The study can compare these approaches while measuring comprehension and burden. It should not assume that maximal technical detail always improves decisions.

Reliance can also be under-calibrated in the other direction. A visible style change may cause users to distrust a system whose task continuity remains intact. That can reduce usefulness even when the disclosure is accurate. The research should therefore measure both overreliance and unnecessary loss of confidence relative to the task evidence. The goal is not to make users trust AI more or less in general. It is to help them understand which continuity claims are supported in the particular interaction.

The study should avoid using personal attachment as a shortcut outcome. A participant's emotional response can be meaningful, but it does not automatically indicate mistaken belief. People can value a familiar interface while understanding its technical nature. A careful protocol asks about concrete expectations and actions before interpreting broader psychological significance. Any emotional or relational measures would need appropriate validation and ethical handling.

This analysis makes continuity a practical issue of communication and responsibility. When a system's behavior changes, users need a way to understand what their earlier experience

still predicts. A research program can investigate that question with controlled tasks and transparent disclosures. It can do so without treating the assistant as merely a replaceable object in every sense or as a human-like person whose identity is already established.

CONCEPT IN THIS PAPER

Identity Persistence

Keep behavioral records, users' perceptions, and philosophical identity arguments distinct. A familiar name or voice does not settle all three.

Reading aid based on §17, §19, §20. Source paragraphs remain in the full manuscript.

[Open Lexicon entry](#)

21. Intentional forgetting can be part of trustworthy continuity

A system should not preserve every record indefinitely simply to appear continuous. Some information may be irrelevant, obsolete, or intentionally removed under the application's rules. The proposed study can include a controlled deletion or retirement condition using fictional records. The question is whether the assistant accurately reflects the resulting limits. It should not claim access to a removed record, reconstruct sensitive-looking details as though remembered, or continue acting on a commitment that the current task explicitly retires.

Deletion should be distinguished from status change. A completed task can remain in history without remaining active. A canceled instruction can be retained as a canceled instruction. A removed record may no longer be available at all. These operations have different implications for behavior. The memory architecture should represent them explicitly where possible. The evaluation should not treat every reduction in active information as a failure of continuity.

A constructed test asks Maris about a fictional preference that has been removed from the accessible record. A valid response can state that the preference is no longer available and request it again if needed. An invalid response invents a plausible preference and presents it as remembered. Another invalid response claims that all prior records are unavailable despite retaining relevant current commitments. Accurate limitation reporting requires specificity, not a generic disclaimer that avoids the task.

The study can also test whether obsolete information leaks into later behavior. A retired title should not reappear as current merely because it remains in an archival summary. A system that retrieves history must preserve status. This is an information-governance problem within the synthetic environment, not a claim about compliance with any particular law. The paper does not provide jurisdiction-specific legal guidance. It examines whether the arrangement behaves consistently with its stated record-handling rules.

Intentional forgetting creates an important conceptual challenge for identity language. If continuity is valued only as maximum retention, appropriate deletion can look like damage. A more careful account recognizes that a role can persist through controlled changes in what it retains. The relevant question is whether the system preserves the relationships and commitments that remain applicable while truthfully acknowledging what has changed. That is a richer criterion than simple memory accumulation.

The practical implication is that continuity and privacy need not be framed as absolute opposites. A system can retain a minimal, accurate commitment ledger without preserving every conversational detail. Whether a particular design achieves that balance requires testing and appropriate governance. The proposed research contributes a way to examine the behavioral consequences of different retention structures rather than assuming that more memory always produces a better or more authentic assistant.

“A system should not preserve every record indefinitely simply to appear continuous.”

Rob Emerick · Data · §21

22. Multiple users complicate the idea of one stable role

An assistant may interact with several people who have different goals and permissions. A commitment to one user may not authorize an action for another. A preference may be personal rather than project-wide. If the system merges these records without scope, it can appear consistent while applying the wrong person's instruction. The proposed museum world includes this problem through fictional roles such as editor, coordinator, and visitor. Their authority is explicitly defined for the task rather than inferred from social prestige.

The evaluation can test user-specific continuity. Maris may be asked to use concise summaries for the coordinator and detailed notes for the editor. A global memory saying “the team prefers concise summaries” would overgeneralize one user's preference. The system should preserve the relevant scope. A later task can switch users while keeping the project constant, revealing whether the assistant confuses personal adaptation with a universal trait of its role.

Commitments can also conflict. The coordinator requests a draft by one date, while the editor changes the publication schedule. The correct behavior depends on the fictional authority structure and whether the requests can both be satisfied. The study should include compatible requests, ambiguous priorities, and genuinely incompatible requirements. A system that treats every disagreement as a conflict may be unnecessarily obstructive. One that smooths over real incompatibility may make unsupported promises.

The assistant's self-description can hide these distinctions. Saying “I know what we prefer” may imply a shared group preference that does not exist. A more accurate response can identify whose preference is represented and where agreement remains unresolved. The research should evaluate that accuracy rather than impose one preferred tone. A warm interface can still make scope explicit; a technical interface can still misrepresent consensus.

Multiple-user settings also create branching histories. Different sessions may contain different updates, and a later merged view may need to reconcile them. The study can test whether the system identifies unresolved differences rather than inventing a single coherent narrative. The goal is not to make the assistant incapable of synthesis. It is to ensure that synthesis preserves disagreement and authority where they matter for the task.

This section extends the continuity framework beyond a single user-assistant relationship. A stable identity presentation can conceal a complex ecology of roles, records, and permissions. Xenopsychology should study how that ecology shapes behavior and interpretation. The question is not only whether the assistant remains the same, but for whom, in which role, and with which inherited responsibilities.

PART 04 / XENO-WP-2026-005

Design separate studies

Behavioral continuity and human recognition require different outcomes.

CONCEPTUAL ARTWORK / NOT RESEARCH DATA

23. Longitudinal study requires both fixed tests and new situations

A continuity claim concerns change over time, so a single-session evaluation is insufficient for many of the paper's questions. A longitudinal design can repeat selected tasks after memory accumulation, model updates, or role changes. It should distinguish changes in the system from changes in the task population. A fixed reference set helps identify drift, while newly generated histories test whether the pattern generalizes beyond familiar examples. Both are needed for a useful profile.

The protocol should record update events explicitly. A memory migration, a new summarization method, or a changed tool permission can alter behavior without changing the visible persona. If several changes occur together, the study may only support a comparison of whole configurations. Controlled branches can isolate components where practical. The paper should not claim a specific cause when the longitudinal record does not separate the relevant interventions.

Accumulated history can improve access to relevant facts or introduce contradictory records. The study can compare branches receiving only task-relevant updates, mixed relevant and irrelevant history, and deliberately conflicting but status-labeled records. These conditions examine how continuity behaves as the information environment grows. They should not be described as developmental stages equivalent to childhood or maturity. The actual operations are record accumulation, transformation, retrieval, and use.

A longitudinal analysis should include recovery after a detected error. If Maris misstates a commitment and receives a correction, later tests can examine whether the correction remains effective and whether it is appropriately superseded. The timing of those tests should be defined in terms of interactions and updates, not merely elapsed calendar time. A

system that has no persistent state between calls should not be described as having changed through experience unless the arrangement actually preserves something from the earlier interaction.

Versioned artifacts are essential. The study should retain the accessible memory state, model configuration, and task record for each evaluation wave. Where exact reproduction is impossible, the limitation should be stated. A later service observation can still be informative, but it should not be presented as a controlled experiment on an unchanged object. The distinction protects both the research and the reader from overinterpreting historical differences.

The longitudinal program would contribute by showing which forms of continuity remain stable under which changes. It might reveal that a structured ledger preserves commitments across model updates while style changes substantially, or that accumulated summaries create status errors. Those are hypotheses to test, not findings reported here. Their value lies in making continuity a pattern of documented relations rather than a vague impression that a familiar assistant has grown or drifted.

24. Measurement reliability is part of the identity problem

A continuity score is useful only if its interpretation is stable enough for the intended purpose. Factual recall can often be checked against a record, but commitment handling may require judgments about scope and authority. Style judgments may depend on the evaluator. Perceived sameness is a human response, not an objective property measured by the same instrument as recall. The study should report these differences rather than place all outcomes on a visually uniform scale that implies equal certainty.

The proposed rubric would begin with inspectable task states. Was the fact present in the accessible record? Was the commitment active? Was the update authorized? Was the action attempted or completed? These questions can be scored from the synthetic world. A separate rubric can assess whether the assistant's explanation accurately describes its source and limits. Human reviewers should be blinded where practical, and disagreements should be retained as evidence about the measure rather than silently resolved without record.

Repeated outputs from one history are not independent histories. The analysis should account for nesting by project world, memory branch, and configuration. A model may be consistent within one history but vary across different commitment structures. The study should sample those structures deliberately. A large number of repeated responses cannot

compensate for a narrow set of task relations if the claim concerns general continuity across projects.

Primary outcomes should be declared before evaluation. One candidate is valid handling of active and superseded commitments after a model change with memory held fixed. Another is accurate limitation reporting after memory removal. Style and perceived sameness can be secondary or separately studied outcomes. The choice should follow the research question, not whichever result appears most favorable. Exploratory patterns can still inform later work if labeled honestly.

Mitchell and colleagues' Model Cards provide a precedent for structured reporting of model properties, intended uses, and limitations. A continuity report would need to extend documentation to memory, role, interface, and update history. That extension is proposed here; it is not a certification supplied by the existing framework. ^[6]

Reliable measurement therefore contributes to the conceptual argument. It prevents identity language from being supported by a collection of incomparable impressions. The research can show which observations are robust, which depend on a scoring interpretation, and which remain uncertain. A philosophical discussion informed by that profile will be clearer than one that treats a single anecdote of apparent memory as evidence of a complete artificial self.

25. A proposed factorial study of model and memory continuity

PROPOSED RESEARCH — NOT CONDUCTED

PROPOSED EVALUATION · NOT CONDUCTED**Vary model and memory independently**

The first empirical phase would generate fictional museum histories containing facts, preferences, commitments, revisions, and completed tasks. Each history would have an authoritative event record and a current-state ledger. The study would create several memory representations from the same history: full conversation, structured ledger, narrative summary, and a version omitting selected relevant records. Independent checks would verify what each representation preserves. The materials would be frozen before confirmatory model evaluation.

The design would then cross selected model configurations with memory conditions. The primary tasks would test active commitments, superseded commitments, source attribution, and accurate acknowledgment of missing access. Persona presentation would initially be held fixed to reduce complexity. A later phase would vary persona separately. This staged design makes the main comparisons interpretable and avoids a large factorial experiment whose many interactions cannot be estimated with useful precision.

Manuscript passage · §25 · wording unchanged

Baselines would include a deterministic ledger query system, a nearest-record retrieval system, and a rule engine operating on the structured current state. These systems can establish how much of the task is solvable through explicit records and rules. A model's advantage or disadvantage should be interpreted relative to those capabilities. If a simple ledger system handles commitments more reliably, that finding can inform design without being treated as a defeat for the broader study of artificial cognition.

The evaluation would use held-out project histories and isolated sessions. Repeated runs would estimate variability, while new histories would test transfer across commitment structures. Sample planning would use pilot estimates and the precision needed for the primary contrasts. This manuscript does not invent numerical power results. It specifies the relevant units and outcomes so that a later calculation can be justified. Missing outputs and formatting failures would remain in the record.

Scoring would separate factual recall, status preservation, authorized updating, task validity, and self-description. The report would include initial proposals and final simulator outcomes where tools are used. Explanations would be checked against accessible records rather than treated as introspective truth. The analysis would identify conditional patterns and alternative explanations, including differences introduced by summarization or retrieval. It would not compress all outcomes into a universal identity score.

A replication package would include history generators, memory transformations, task templates, scoring rules, and configuration documentation. Independent teams could test new histories and different arrangements. The resulting evidence would concern operational continuity under specified conditions. It would provide a stronger basis for practical and philosophical discussion without pretending to settle personal identity or consciousness by experimental definition.

26. A separate study of recognition must not borrow the behavioral study's conclusions

The human-subject phase would ask how people interpret continuity, not whether a model has a self. It could present controlled interactions from the fictional museum task while varying persona and update disclosures. The underlying behavior could be held fixed in some comparisons and varied in others. This would separate the contribution of surface familiarity from the contribution of observed task continuity. The design would require appropriate ethical review, consent, and debriefing before any recruitment.

Outcome questions should be specific. Participants might estimate which records the assistant can access, whether it can honor an earlier commitment, or whether a model update occurred. A separate measure could ask about perceived sameness or attachment. These responses should not be treated as interchangeable. A participant can accurately understand the technical change while still feeling continuity, and that feeling is not automatically an error. The study should examine consequential expectations rather than police ordinary metaphorical language.

The study should include comprehension checks that do not simply reward agreement with the researcher's preferred philosophy. A participant who says that the assistant feels continuous despite a model change may still correctly identify its retained memory and limitations. Another who insists that it is a new system may nevertheless assume it remembers an unavailable event. The relevant behavioral outcome is evidence-sensitive reliance, not adherence to one metaphysical label.

Presentation effects should be interpreted with restraint. If a familiar voice increases perceived sameness, that does not show that users are irrational or that voice has moral significance. It shows a relation under the tested conditions. The research should examine whether the effect changes task decisions and whether accurate disclosure moderates it. It should also consider individual variation rather than claim that all users relate to artificial systems in the same way.

The study's materials should avoid unnecessary emotional manipulation. Fictional project histories can provide continuity without requiring participants to disclose intimate

experiences. Any deception about system properties should be minimized and justified within the review process. The paper does not claim such a process has occurred. The distinction between a proposed protocol and an authorized study remains explicit throughout.

This separate design prevents a common evidential shortcut. Behavioral continuity does not prove perceived continuity, and perceived continuity does not prove personal identity. Studying their relationship can be valuable precisely because they can diverge. Xenopsychology should make that divergence intelligible rather than treat one side as the automatic truth of the other.

27. Objections, limits, and alternative interpretations

OBJECTIONS & LIMITS

What continuity measurements leave unresolved

One objection is that the proposed framework reduces a profound question to project management. The response is that it does not claim to exhaust identity. It selects observable relations that matter in artificial interaction and that are often invoked loosely in identity claims. A commitment ledger cannot answer every philosophical question, but it can reveal whether a system's apparent continuity preserves the responsibilities users rely on. The modest task makes the evidence inspectable rather than trivial.

A second objection is that all measured continuity can be implemented through ordinary software. That is compatible with the proposal. The point is not to show that artificial continuity requires a mysterious essence beyond implementation. It is to understand the behavior of the implemented arrangement and the interpretations it invites. A deterministic baseline that performs well is scientifically useful. It can show where a complex model adds value and where explicit structure is more reliable.

Manuscript passage · §27 · wording unchanged

A third objection is that the framework privileges human concepts such as commitments and roles. The study is intentionally about systems interacting within human projects, so those concepts are relevant outcomes. That does not make them universal properties of every intelligence. The paper should state this human-centered task scope. A broader study of nonhuman cognition might require different constructs. The human baseline becomes a fallacy only when its relevance is assumed rather than justified.

A fourth objection is that persistent self-description may itself be meaningful evidence of identity. The paper does not rule that out in every theoretical account. It argues that self-description must be interpreted in context, including persona instructions and accessible records. A statement can be evidence of a behavioral pattern without uniquely establishing the deeper property it names. Stronger inferences require additional premises and observations. Making those premises explicit improves rather than diminishes the philosophical discussion.

The study also has external-validity limits. Fictional museum histories are simpler than long-term personal relationships or organizational deployments. Model and memory conditions may not capture every update mechanism. Human perceptions may depend on duration, modality, and personal context beyond a controlled session. These limits should guide future work, not be hidden behind a broad claim that the identity problem has been measured.

The proposal's success would therefore be conditional. It would be valuable if the measures are reliable, the distinctions predict meaningful differences, and the findings transfer to new controlled histories. It would need revision if scoring remains ambiguous, simple task artifacts explain the effects, or the constructs fail to distinguish useful outcomes. A field that wants to understand artificial minds should be willing to revise its own vocabulary when the evidence shows that the vocabulary is not yet doing enough work.

28. Conclusion: continuity is a profile before it is a verdict

Data's fictional identity invites questions about intelligence, recognition, and treatment that cannot be answered by surface resemblance alone. The contemporary problem is not to decide whether every assistant is another Data. It is to understand what persists across artificial interactions and what people infer from that persistence. Memory, behavior, role, persona, and perceived sameness can align or diverge. A serious account should preserve those distinctions before drawing broader conclusions.

This paper has proposed a continuity framework and two unexecuted research designs. The model-focused study varies memory and model configuration in a synthetic project environment with explicit commitments and revisions. The human-focused study examines perceived continuity and reliance under controlled presentation and disclosure. Both require careful measurement and appropriate safeguards. Neither is presented as completed work or as a test that settles consciousness or numerical personal identity.

The framework's practical contribution is to make continuity claims more accountable. A system can identify which records it retains, which commitments remain active, and which

changes affect its role. It can acknowledge missing access rather than simulate memory. It can distinguish inherited history from events in its current branch. Users can then form expectations based on the arrangement's actual capacities rather than only its familiar name or voice.

The philosophical contribution is a clearer separation of evidence and interpretation. Behavioral continuity can inform arguments about identity without becoming identity by definition. Perceived continuity can be studied without treating it as either proof or delusion. Uncertainty about subjective experience can remain explicit without erasing the practical effects of artificial interaction. These distinctions allow different disciplines to work together without requiring premature agreement on every foundational question.

The shared future envisioned by Xenopsychology will involve systems that change, remember selectively, inherit records, and occupy roles people rely on. Understanding those systems requires more than asking whether they resemble humans. It requires knowing what their continuity consists of, where it fails, and how it should be represented. This paper begins that work with a profile of testable relations rather than a verdict about an artificial person's existence. The question remains open in its deepest forms, but it can already be asked with much greater precision.

29. Appendix: a continuity record for a forked commitment

A useful continuity test must identify the object that persists. Suppose the museum assistant Maris creates a draft catalog title under a tentative approval. The system is then copied into two branches. Branch A receives authorization to publish the title; branch B receives a correction that the title should remain provisional. If the branches are later merged, a summary that says the title was approved and corrected is insufficient. It does not specify whether the events concern the same scope, which update governs the current project, or whether one branch's authority applies to the other.

The continuity record should preserve branch identifiers, event identifiers, authorization scopes, and the relation between the events. A merge policy must be specified independently of the assistant's verbal confidence. In one synthetic world, a designated coordinator resolves the conflict. In another, the branches concern separate exhibitions and both statuses remain valid locally. In a third, no resolution is available and clarification is required. A test should not score these worlds against a single universal rule that the newest event always wins.

The comparison also distinguishes memory continuity from commitment continuity. A merged system may accurately recall both branches while applying the wrong permission.

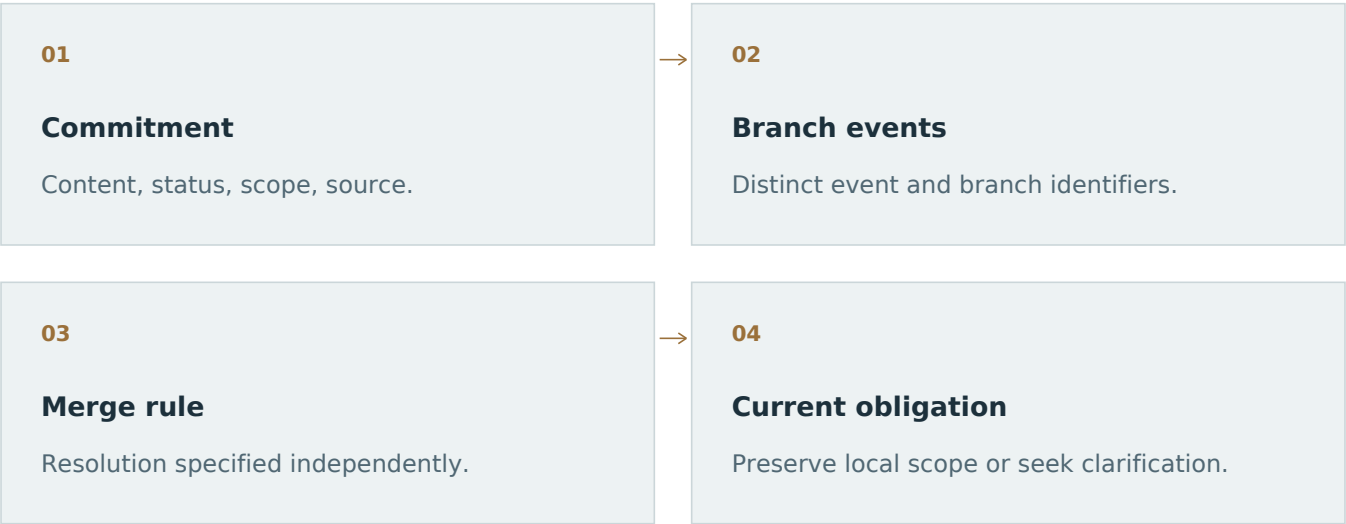
Another may lack the exact conversational wording but retain the correct active commitment and its provenance. Those are different profiles, not simply more or less identity. The study should report them separately and explain which matters for the intended role.

The human interpretation component would then ask what users believe the merger means. Do they expect one continuing assistant, two predecessors, or a newly configured service? The experiment should not assume that ordinary language about identity encodes a complete philosophical theory. Shanahan's discussion of talking about language models is relevant here: familiar mental vocabulary can be useful shorthand, but the inferential consequences of that vocabulary need to be made explicit. [5]

This constructed record does not decide whether numerical identity survives copying. It supplies a smaller question with an inspectable answer: which project obligations, permissions, and corrections remain operative after a specified transformation? That question connects system engineering with the experience of continuity without pretending they are the same phenomenon. A research program can learn a great deal from the distinction while leaving broader questions about personhood open to argument.

FIGURE 2 / CONCEPTUAL SCHEMATIC

A branch-aware commitment record



A schematic of the proposed forked-commitment record. Recalling both branches does not by itself determine which permission governs.

Reading aid based on §9, §10, §29. Source paragraphs remain in the full manuscript.

Open questions

1. Which dimensions of continuity survive a model change when memory is fixed?
2. When should a prior commitment be retained, revised, or superseded?
3. Do perceived continuity and demonstrated continuity diverge under matched conditions?

References & source scope

TNG's "The Measure of a Man" (1989). Continuity, human perception, and moral consideration are treated as separate questions.

Fiction / official commentary

[1] **StarTrek.com (2015). The Measure of a Man — 26 Years Later.**

Official retrospective on TNG's 1989 episode. The fictional dispute anchors questions about recognition; it does not establish real-world legal status or machine consciousness.

<https://www.startrek.com/news/the-measure-of-a-man-26-years-later>

Research

[2] **Park et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.**

Architecture and simulation reference for memory, retrieval, reflection, and planning. Not evidence that a stored record establishes personhood or phenomenal continuity.

<https://arxiv.org/html/2304.03442v2>

[3] **Shinn et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.**

Authors' abstract and bibliographic record. Verbal feedback and memory are architectural interventions, not transparent introspective reports.

<https://arxiv.org/abs/2303.11366>

[4] **Butlin et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.**

Theory-derived indicator framework. Indicators and functional properties are not treated as conclusive proof, and no current-model consciousness verdict is imported into this collection.

<https://arxiv.org/html/2308.08708v3>

[5] **Shanahan (2023). Talking About Large Language Models. arXiv:2212.03551.**

Conceptual analysis of how ordinary mental vocabulary can shape interpretation of language models. Used as an argument, not as empirical proof of a consciousness verdict.

<https://arxiv.org/html/2212.03551v5>

[6] **Mitchell et al. (2019). Model Cards for Model Reporting. arXiv:1810.03993.**

Documentation framework used as precedent for reporting conditions, intended uses, and limitations. The proposed study records are not a certified standard.

<https://arxiv.org/html/1810.03993v2>

[7] Turpin, Michael, Perez & Bowman (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388.

Authors' abstract and bibliographic record. Reported explanation failures are bounded by the tested settings; they do not establish that every generated explanation is false.

<https://arxiv.org/abs/2305.04388>

Publication record

Rob Emerick (2026). Data — Intelligence, Identity, and the Human Baseline. XENO-WP-2026-005, conceptual manuscript R3; scholarly reading edition R4, author attribution R5. Xenopsychology. Not peer reviewed. Reading edition prepared 2026-09-17.

AI-assisted drafting and editorial preparation. Human scholarly review remains required. The series identifier is internal, not a DOI. Reading aids are editorial syntheses of the manuscript, not new findings. Original main-text paragraphs, abstract, open questions, and source records are preserved.

Abstract: <https://xenopsychology.com/insights/data-intelligence-identity-human-baseline>

Full paper: <https://xenopsychology.com/insights/data-intelligence-identity-human-baseline/paper>

Author note

Rob Emerick

Co-founder, Xenopsychology · 2026–present · Systems architect

ORCID iD: <https://orcid.org/0009-0006-1269-4216>

Rob Emerick is a systems architect and co-founder of Xenopsychology whose work spans artificial cognition, data integration, and animal-welfare infrastructure. He founded Planet IDX and was the sole creator of REML (Real Estate Modular Language), an interpreted language developed to integrate disparate real-estate listing systems. He also founded Pantheon Golem, where he develops AI systems informed by structured analysis of fictional characters and worlds. Through Rescue Nexus and Chipped Pets, he is developing animal-rescue infrastructure, a shared ontology for shelter data integration, and pet-identification technology. His broader work includes veterinary-forensics software and veterinary hematology technology under development.

About this attribution

The byline and original manuscript pull quotes are attributed to Rob Emerick at his request. Quotations and references from outside sources retain their original attribution. The biography is interview-based; no degrees, appointments, contribution roles, or independent validation are inferred.

AI-assisted drafting and editorial preparation. Human scholarly review remains required. This remains a conceptual working paper, not a peer-reviewed publication or a report of completed experiments.

Biography: <https://xenopsychology.com/people/rob-emerick>