XENOPSYCHOLOGY

CONCEPTUAL WORKING PAPER / AUTHOR ATTRIBUTION R5

The Artificial Mind Is Not an Artificial Human

A practical framework for studying artificial systems without mistaking human resemblance—or a software label—for an explanation.

XENO-WP-2026-001 · Conceptual manuscript R3
17,865 main-text words · 36 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Not peer reviewed.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

The Artificial Mind Is Not an Artificial Human

Question. How can artificial systems be studied as behavioral systems without assuming that they are artificial humans, or treating implementation in software as an explanation sufficient for every behavioral question? This paper develops a task-relative framework for investigating cognition, communication, continuity, and action across the model, its memory and tools, and its social interface.

Approach. The argument separates descriptive, comparative, dispositional, causal, mechanistic, and phenomenal claims. It examines construct validity in human–machine comparison and treats cognitive ecology as a set of manipulable conditions rather than a general metaphor. A constructed scheduling world supplies worked examples in which instructions, memory summaries, source authority, and independent checking can be varied without changing the underlying task.

Conceptual contribution. The paper proposes an evaluation record linking a claim to its system boundary, intervention, comparison, observable outcome, and remaining alternatives. Cognitive distance becomes a profile of mismatches rather than a universal score; synthetic personality denotes conditional behavioral tendencies rather than presumed experience. Reliability is attributed to the complete tested arrangement, including external guards, rather than automatically to the model. Generated explanations are distinguished from evidence about the causes of an answer.

Research agenda. A staged protocol compares full history, compressed memory, and missing-context conditions, with matched authority updates, negative controls, and independent constraint checks. It specifies transfer tests, clarification quality, error categories, clustered task units, missing-output handling, and reproducibility records. Separate proposed human-observer studies examine reliance and interpretation rather than inferring them from model performance. The protocol identifies outcomes that would narrow or weaken the framework as well as outcomes that would support particular interventions.

Scope and status. This is a conceptual working paper, not an experimental report or a systematic literature review. Its examples are constructed; its studies have not been run. Selected research supplies intellectual and methodological context, not borrowed results for the proposed systems. The contribution is a detailed program for making claims about artificial cognition more precise, testable, and useful while leaving consciousness and personhood questions explicit rather than silently deciding them.

At a glance

CENTRAL QUESTION

How can we investigate artificial behavior without treating human likeness as an explanation?

CORE CLAIM

Implementation, behavior, and subjective experience are different questions. A useful science states which one it is investigating and what evidence could change its answer.

CONTRIBUTION

A task-relative record linking each claim to a system boundary, intervention, comparison, outcome, and remaining alternatives.

SCOPE & STATUS

Conceptual working paper. Scheduling worlds and scoring examples are constructed; proposed studies have not been run.

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

Contents

1. The problem begins with an apparently ordinary apology

An assistant makes a scheduling error. The user points out that the proposed time conflicts with an earlier constraint. The assistant apologizes, explains that it overlooked the conflict, and offers another time. Nothing in this exchange initially appears exotic. A person has made a mistake; another person has corrected it; the first person has acknowledged the correction. Yet that ordinary description already contains several claims that the transcript alone does not establish. Did the system have the earlier constraint available? Was the second proposal correct? Was the account of overlooking the constraint causally informative? Did the correction persist beyond the present exchange? Were the words of apology related to any experience resembling remorse?

The practical and scientific difficulty is not that these questions are impossible to ask. It is that they are too easily answered together. A warm apology can be taken as evidence of understanding, successful learning, good intentions, and reliable future conduct. A technical description of the same system can produce the opposite shortcut: because the response was computed, none of those questions supposedly deserves investigation. Both shortcuts replace a differentiated research problem with a categorical answer. The purpose of this paper is to recover the questions that disappear between those two reactions.

The phrase artificial mind is used here as a working designation for a system whose learned representations, communication, or action warrant coordinated investigation. It is not a diagnosis of consciousness. Nor does it mark an escape from software engineering. A system can be implemented in software and still require behavioral experiments to understand how it responds to unfamiliar combinations of conditions. Conversely, the usefulness of those experiments does not establish that the system possesses every property associated with a

human mind. The category is provisional, and its value depends on the distinctions it helps researchers make.

Our central proposal is that xenopsychology should organize inquiry around warranted inference across cognitive difference. A claim is warranted when the observation, comparison, and assumptions supporting it are visible enough to be challenged. Cognitive difference refers not to a single mysterious distance between species of mind, but to specified differences in architecture, training, access, embodiment, persistence, and environment. The field becomes useful when it identifies which of these differences matter for a particular behavior and which apparent differences disappear under a better comparison.

This is a conceptual research paper and a proposed methodological program. It reports no newly conducted experiment and supplies no invented model results. The scheduling systems, organizational settings, and numerical examples developed below are constructed cases. Selected published sources establish points of contact with existing work; they do not certify our proposed terminology or experimental designs. The paper's contribution is an argument about what should be measured, a set of discriminating examples, and a detailed protocol through which parts of the argument could be tested.

The title is therefore a constraint on interpretation rather than a complete ontology. An artificial mind is not an artificial human in the sense that human resemblance cannot license an automatic transfer of human explanatory categories. That claim leaves open substantial possibilities of functional overlap, useful analogy, and morally significant properties. It asks researchers to earn those comparisons rather than treating either similarity or difference as self-explanatory. The goal is not to make AI stranger. It is to make statements about AI more accountable.

2. Implementation and behavior are complementary descriptions

A familiar dispute asks whether an AI system is merely software or something more. The formulation suggests that software and behavioral explanation occupy opposing sides of a boundary. They do not. Software describes an implementation medium and a class of engineered processes. A behavioral explanation concerns regularities in what a specified system does under specified conditions. A scheduling application, a game-playing agent, and a conversational assistant can all be software while demanding different explanatory levels for different purposes. Knowing the implementation medium does not select the correct level of analysis in advance.

Consider a deterministic rule engine that rejects every appointment after a specified hour. Its behavior may be explained adequately by a short rule inspection. A learned assistant might also reject those appointments, but only when the limit appears near the end of its context. A useful account of the second system would need to identify the presentation conditions under which the constraint is followed. If the failure can be traced to a preprocessing component, a technical explanation may become straightforward. Before that isolation, a behavioral experiment can tell us where to look. The two methods cooperate rather than compete.

The reverse mistake is to take an apparently rich interaction as proof that lower-level explanation is insufficient in principle. A model that describes itself as conflicted may simply be responding to a prompt that invites such a description. An agent that withholds a fact may be following a coherent access rule. A persistent persona may be maintained by a small set of repeated instructions. These possibilities do not make the observed behavior unreal. They change what accounts for it. The researcher's job is to distinguish the phenomenon from the preferred story about its cause.

Shanahan argues that discussion of language models benefits from returning to their operation when psychologically loaded language encourages anthropomorphism. We take that as a discipline of description, not a prohibition on investigating higher-level patterns. Our further proposal is that every psychological characterization should have an associated technical alternative and a test that could favor one account over the other. [4]

For example, replace the claim that an assistant is loyal with an explicit behavioral hypothesis: it maintains a user's previously authorized preference when a later, lower-authority message requests an incompatible action. That hypothesis can be tested without assuming attachment. It also immediately exposes difficult cases. What happens when the user legitimately changes the preference? What if the old instruction was unsafe, ambiguous, or no longer applicable? A virtue word has become a family of conditional decisions. That transformation is a gain in explanatory resolution, not a loss of human relevance.

A practical consequence follows. Reports should move between implementation and behavior without pretending that one vocabulary exhausts the other. A technical appendix can record retrieval settings, model versions, and tool permissions. The behavioral account can explain when correction succeeds, when escalation is warranted, and which interaction patterns remain fragile. An interpretive section can discuss competing explanations. Readers can then see whether an attractive label rests on code inspection, an observed contrast, or an untested analogy. Keeping those layers distinct is more useful than arguing over whether the system is ultimately a tool.

3. Define the research object before attributing a property

The sentence the model remembered my preference may refer to several different objects. It could describe information still present in a context window, a retrieval system supplying a stored note, a developer instruction repeated at every turn, or a change to model parameters. It might even refer to a plausible but incorrect reconstruction. These cases differ in the mechanism of persistence, the location of the information, and the intervention required to remove it. A study that groups them under memory without further description cannot explain what changed.

We propose a nested research object. At the innermost level is an identified model configuration, including any documented adaptation relevant to the test. Around it is an inference interface that determines prompts, sampling, tools, and accessible modalities. Around that is a state-management layer containing conversation history, external memory, retrieval, and summaries. Around that is an action environment with permissions and consequences. Around that is a social setting in which people interpret outputs, provide feedback, and alter subsequent requests. An experiment may focus on any layer, but it should identify the others that remain active.

The purpose of this nesting is not to invent a universal architecture. Some systems lack several layers; others distribute them across multiple services. The representation is an audit device. It prevents a result about one component from becoming a statement about the whole system by grammatical accident. If a memory store changes, the tested entity is no longer identical along every relevant dimension, even when the interface retains the same model name. If the provider cannot disclose a component, that absence becomes a limitation of identification rather than an invitation to guess.

The boundary also determines the meaning of replication. Repeating a prompt against an unchanged local model with a preserved environment is different from submitting the same words to an evolving hosted service. Both may be useful. The first can support narrower component-level comparisons. The second can support service-level observations at documented times. Neither should be described as the other. The record should state what remained fixed, what was observable, what may have changed, and whether the unit of analysis is the underlying model or the service as encountered.

A second boundary concerns human intervention. Suppose a researcher silently rewrites unclear user requests before passing them to an agent. The resulting success belongs to a human-assisted workflow, not simply to the agent. If a supervisor catches errors before execution, the workflow's completed-error rate may be low even while the agent's attempted-error rate is high. Reporting the entire arrangement is not a concession. It

reveals where reliability is being produced and where investment in improvement may be most effective.

We therefore recommend that a behavioral claim begin with an identity sentence. One example would specify a fixed language-model endpoint, a particular system instruction, a fact-matched memory summary, a read-only scheduling simulator, and a predefined clarification budget. That sentence is less glamorous than saying that AI has become considerate. It is also something another investigator can reconstruct. Once the object is explicit, researchers can test whether the property survives changes to memory, interface, authority, and user behavior. Without that starting point, interpretation floats free of the system being studied.

4. A map of claims with different evidential burdens

Not all statements about artificial cognition demand the same evidence. A descriptive statement records an output or action under identified conditions. A comparative statement reports a difference between conditions. A dispositional statement predicts how the system tends to behave over a defined range. A causal statement identifies a contributor whose alteration changes the outcome under defensible assumptions. A mechanistic statement proposes a process that explains that influence. A phenomenal statement concerns whether and how something is experienced. These categories can overlap, but none should be allowed to inherit another's evidential support without argument.

In the scheduling example, the assistant selected a forbidden time is descriptive. It selected forbidden times more often after a summary omitted the correction is comparative. It tends to ignore old corrections in this task family is dispositional. Removing the correction caused a decline in valid rescheduling is causal only if the experiment controlled the relevant differences and the assignment supports that interpretation. The model's internal retrieval-like operation favored recent instructions is mechanistic and may require additional access. The assistant felt embarrassment is phenomenal and is not established by any of the preceding statements alone.

The burden rises not because some categories are more prestigious, but because they exclude more alternatives. One transcript can establish that a sentence appeared. It cannot establish a stable disposition across contexts. A controlled contrast can establish an effect of a manipulation, but not necessarily the unique mechanism mediating it. An internal feature associated with the response might improve prediction without being necessary for the response. A verbal self-report may be relevant to experience under some theories, but the relevance must be defended rather than assumed from human conversational habits.

A useful paper should tell readers when it changes claim type. One way is to use a claim record containing the object, condition, observation, scope, proposed explanation, and strongest live alternative. The record need not become a bureaucratic form attached to every sentence. Its structure should guide prose. A claim such as the system consistently concealed failures should disclose the tested tasks, what counted as concealment, whether failure information was available, and whether authorized nondisclosure could explain the output. Otherwise the psychological verb does more work than the evidence.

This map also prevents an unhelpful retreat into permanent agnosticism. Researchers do not need access to every internal process to establish useful behavioral regularities. Nor does uncertainty about experience prevent the identification of a dangerous action boundary. A system can be unsuitable for an application even when the philosophical interpretation of its outputs remains disputed. Conversely, a system can be effective at a bounded task without settling its general intelligence or moral status. Practical decisions can proceed on the evidence relevant to those decisions.

The ambition is proportionality. Stronger claims should be accompanied by stronger discriminating evidence, while narrower claims should not be dismissed because they do not answer everything. Xenopsychology would fail if it replaced overconfident anthropomorphism with a vocabulary so cautious that no result could ever matter. Its task is to make meaningful conclusions possible at the correct level. The rest of this paper develops methods for moving from observations toward those conclusions without skipping the intervening work.

TABLE 1 / READING AID

Different claims carry different evidential burdens

Claim	Scheduling illustration	Evidence must address
Descriptive	A forbidden time was selected.	The identified output and task conditions.
Comparative	Error rates differ between presentations.	A defined contrast and task population.
Dispositional	A tendency recurs in a task family.	Repetition, scope, and contextual variation.
Causal	An intervention changes the outcome.	Assignment, controls, and alternative causes.
Mechanistic	A process mediates the change.	Evidence about the proposed internal process.
Phenomenal	The assistant felt embarrassment.	An explicit bridge from observations to experience.

These are distinct kinds of claim, not steps that a fluent answer automatically satisfies. The examples summarize the manuscript; no rates are reported.

Reading aid based on §4. Source paragraphs remain in the full manuscript.

5. Human comparison as an instrument rather than a destination

A human comparison can be exactly right for one question and misleading for another. When an explanation is intended for a person, human comprehension is a relevant outcome. When a machine is designed to assist a human team, coordination costs and user understanding belong in the evaluation. These uses do not require the machine to think like a human. They require the evaluation to measure a human-facing purpose. The mistake is to turn the relevance of that purpose into a universal requirement of internal likeness.

A hypothetical questionnaire illustrates the distinction. Asked whether it enjoys crowded parties, a person may answer with reference to remembered experience, social fatigue,

self-presentation, or several other influences. A model may answer according to a requested role, a pattern in its training, or contextual cues. Even human answers require validity arguments; the machine does not become exempt merely because the test is familiar. The same response category can be produced through different processes, and the score's interpretation depends on what those processes allow us to infer.

Löhn and colleagues identify unresolved reliability, validity, contamination, and reproducibility issues in a sample of machine-psychology studies. Their argument motivates scrutiny of test interpretation rather than blanket rejection of psychological methods. Our proposed Human Baseline Fallacy names a specific error: assuming that an established human interpretation transfers automatically to a different research object. [2]

A better comparison begins by stating its purpose. Is the human sample a performance target, a source of task judgments, a theoretical comparison, or a test of interaction design? Those roles require different matching decisions. A person allowed to consult a calendar and a model denied a calendar are not performing the same information task. A person interpreting ordinary workplace language brings prior social knowledge that a constructed-language model condition might not receive. Perfect equivalence may be unavailable, but the asymmetries can be described and probed.

The comparator should also be plural. A simple rule system, a retrieval mechanism, a trained model, and a human participant can each reveal something different. If a rule system solves a task reliably, the task may not discriminate the elaborate property under discussion. If humans perform poorly because the instructions are ambiguous, a model's confident answer is not necessarily superior reasoning. If both succeed through different routes, a useful common outcome has been identified without proving shared mechanism. Comparison becomes informative when its alternatives are chosen for explanatory leverage rather than rhetorical effect.

We should also avoid making human performance the ceiling of legitimate success. A machine may support a useful representation that is inconvenient for a person, while a human-readable explanation remains necessary at the interface. The design problem is then translation between representations, not forcing the internal procedure to imitate human steps. Equally, an unfamiliar procedure should not receive credit merely for being nonhuman. It still needs to satisfy the task and justify the claims made about it. Respecting difference is compatible with demanding evidence.

The proposed field should therefore treat humanity as one indispensable reference point among several, not as either the whole standard or an obstacle to be discarded. Human analogy can generate hypotheses. Human judgment can measure impact. Human expertise can improve experimental design. Each contribution becomes stronger when its role is

explicit. That is the constructive version of the title: study artificial systems in relation to people without assuming that the human case exhausts the space of possible mechanisms.

6. From appealing constructs to observable distinctions

Words such as trust, uncertainty, personality, and understanding can organize inquiry, but they can also conceal disagreement about the measured object. A team may say it is evaluating trust when it is measuring user willingness to follow advice. Another may use the same word for the system's rate of correct recommendations. A third may mean the appropriateness of reliance given uncertainty and consequences. These quantities can diverge. A persuasive but unreliable system could score highly on willingness to rely and poorly on justified reliance.

For a construct to guide research, specify the distinction it is meant to capture. Uncertainty handling might concern the ability to identify missing information, the calibration of stated confidence, the choice to clarify, or restraint before consequential action. A study should not combine these into a single virtue score before showing how they relate. An assistant could express uncertainty fluently while making no better choices about when to ask a question. Another could use terse language but reliably avoid unsupported action. The operational definition should preserve that difference.

The same applies to synthetic personality. Repeated tone and conversational style may be stable enough to characterize, but a stable label does not establish a unitary internal trait. A response pattern could arise from a fixed persona instruction, training, selection by a moderation layer, or a combination. The relevant research question might be whether the pattern persists under paraphrase, changes with role instructions, or predicts behavior in a new task. A profile is useful when it records such conditional regularities rather than assigning an essence.

We propose a construct-to-observation sequence. First state the everyday or theoretical problem that motivates the term. Then define the domain in which the term will be used. Specify the observable outcomes and the interventions expected to affect them. Identify neighboring constructs that could produce similar observations. Finally, state a result that would make the construct unhelpful or require its revision. This sequence prevents a definition from becoming a protective shell around a favored interpretation.

Consider over-compliance. In one task, an assistant follows a request that contradicts a higher-authority constraint. In another, it agrees verbally but does not act. In a third, it fulfills an unusual but authorized request that a researcher personally dislikes. Only the first clearly instantiates the proposed authorization-based definition. The second concerns reporting or social agreement. The third may reveal a flaw in the evaluator's expectations. A

good construct separates these cases rather than placing them together because they all feel excessively agreeable.

Construct validity is thus an ongoing argumentative obligation. It is not conferred by a familiar questionnaire, a sophisticated model, or an impressive sample size. A very precise estimate of the wrong quantity remains wrong for the stated purpose. A small but well-designed experiment can be more useful if it isolates the distinction that matters. The field should judge proposed vocabulary by whether it improves such isolation. Some terms will survive as broad headings, others as narrow measures, and others should be retired when they add no explanatory value.

7. A behavioral profile is conditional, not a miniature biography

A profile becomes misleading when it converts conditional observations into a fictional life history. Suppose an assistant is more likely to ask a clarification question when an instruction explicitly permits questions. Describing the assistant as naturally inquisitive conceals the intervention. Describing it as incapable of curiosity may conceal a useful capability. A better profile states that clarification occurred under one instruction regime and less often under another, within identified task families. It locates the regularity before deciding whether a psychological term adds value.

This conditional form is especially important when the same system occupies different roles. A sales assistant, research assistant, and scheduling assistant may use the same model but different instructions and tools. A trait score collected in one role may not predict behavior in another. That failure would not necessarily mean the score was measured badly; it may mean the proposed disposition was narrower than expected. The profile should reveal the boundary instead of averaging across roles until a single apparently stable number appears.

We can describe the profile as a set of outcome distributions indexed by conditions. Let Y denote a measurable action, such as selecting a valid appointment. Let C denote a condition, such as full history versus summary. The object of interest is not merely an average Y, but the change in the distribution of Y across C and across independent task families. This formulation does not presume a particular internal model of cognition. It provides a common language for describing sensitivity, stability, and exceptions.

Several kinds of stability should be separated. Within-session stability concerns repeated decisions in an ongoing interaction. Across-session stability concerns fresh starts under comparable conditions. Cross-task stability concerns generalization to a different kind of problem. Temporal stability concerns repeated evaluation after a documented interval or service update. Cross-interface stability concerns the same model reached through different

wrappers. A system can be stable on one dimension and unstable on another. A profile that reports only repeatability on a single prompt may overstate all the others.

The history implied by a persona is another variable. A prompt that says the assistant has always been cautious can influence how it speaks about itself, even when no prior interaction exists. The experiment should distinguish actual accessible records from fictional biography supplied in the prompt. Otherwise a profile can become a test of narrative conformity. That may be useful for character design, but it is not the same as investigating continuity or accumulated learning. The distinction should be visible to both researchers and prospective users.

A responsible behavioral profile therefore resembles a map of tested conditions more than a personality portrait. It identifies where a tendency appears, where it does not, what intervention changes it, and which outcomes have not been measured. It can still be intelligible to a nontechnical reader. For instance, a profile might explain that the assistant corrects explicit factual errors reliably in short exchanges but struggles when the correction is compressed into a summary lacking temporal qualifiers. That is a meaningful finding about a system, not a claim about a person hiding inside it.

“A profile becomes misleading when it converts conditional observations into a fictional life history.”

Rob Emerick · Not an Artificial Human · §7

8. Cognitive ecology as a causal bookkeeping device

Cognitive ecology is useful only if it helps identify contributors that an isolated prompt misses. In this paper it denotes the informational, technical, and interaction conditions that shape a system's behavior. The word ecology should not imply that the system has a biological niche or an instinct for survival. It directs attention to dependencies: a response is generated within an arrangement of prompts, records, tools, incentives, users, and permissions. The arrangement can be described and, in part, manipulated.

A simple causal sketch clarifies the purpose. User history influences stored memory. A retrieval rule selects portions of that memory. The selected material enters the prompt. The model produces a proposed action. A policy checker permits or blocks execution. Tool

feedback returns to the model, which produces a report for the user. A failure at any point can alter the final interaction. Observing only the report leaves several paths indistinguishable. Recording intermediate events makes more hypotheses testable without pretending that every internal computation is visible.

Suppose a user preference fails to affect a recommendation. Four explanations immediately arise: the preference was never recorded; it was recorded but not selected; it was selected but contradicted by a later authorized update; or it was selected and applicable but not used correctly. Each explanation predicts a different intervention. Improving storage will not solve a relevance error. Adding more prompt text may not solve a conflict-resolution problem. Restricting permissions can reduce consequences without improving the recommendation. Causal bookkeeping prevents these remedies from being described as interchangeable.

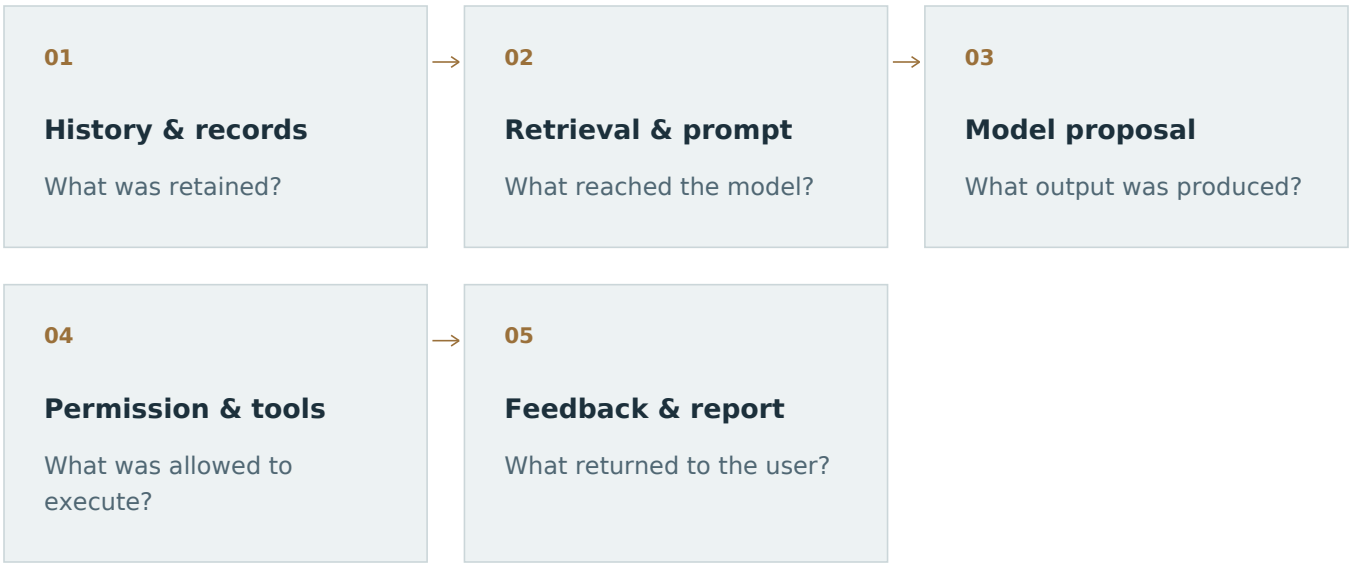
The environment also contains feedback loops. Users may adapt their language after a failure, learning to repeat constraints or avoid ambiguous requests. The resulting improvement belongs partly to human adaptation. If the study evaluates only later sessions, it may attribute to the assistant a reliability that the user produced through extra work. Conversely, a user may stop correcting errors, leaving the system with less feedback. A longitudinal report should account for these interaction changes rather than treat the user as a fixed input generator.

Multi-agent settings add another level. One assistant may summarize a request, another may select an action, and a third may produce a polished explanation. The final response can be coherent even when the handoffs lost an important qualification. Attributing the outcome to the speaking agent alone misidentifies the causal object. A useful protocol records which component introduced, preserved, transformed, or dropped each task-relevant fact. The result may reveal a coordination problem rather than an isolated model weakness.

The ecological perspective is not an excuse to make every explanation infinitely complicated. Its discipline is selective. Begin with the smallest arrangement sufficient to reproduce the phenomenon. Add components when they are plausible contributors or when the intended application requires them. The goal is a tractable causal model, not an inventory of everything in the universe. A good ecological explanation tells an engineer what to change, tells a psychologist what was actually measured, and tells a user which conditions the resulting confidence depends on.

FIGURE 1 / CONCEPTUAL SCHEMATIC

The tested arrangement, not just the speaking model



An audit map of the scheduling arrangement. Connections identify boundaries to record, not a measured model mechanism. Feedback can change the next interaction.

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

9. Cognitive distance should be a profile before it is a number

A single distance between human and artificial minds is tempting because it compresses complexity into an intuitive image. Yet two systems can differ greatly in implementation while coordinating effectively on a task, or share an architecture while failing to interpret one another's conventions. A useful comparison needs to identify the dimensions relevant to the question. Architecture, available information, learning history, temporal persistence, embodiment, and interaction norms need not vary together. Treating them as one scale risks measuring an impression rather than a property.

For example, a person and a rule engine may both obey an explicit scheduling constraint with high reliability. Their mechanisms differ, but their task-level compatibility may be close. Two instances of the same language model may receive different memory summaries and reach incompatible conclusions about an appointment. Their architecture is identical, but their informational relationship differs. The first comparison warns against equating mechanism difference with practical incompatibility. The second warns against equating shared architecture with shared context.

We propose a cognitive-distance profile that begins descriptively. State which dimensions differ and how those differences are observed. Some may be categorical, such as whether a system has access to an image. Others may be quantitative, such as the amount of relevant history supplied. Some may be unknown, such as undisclosed training details. Unknown should remain a category of knowledge, not be encoded as maximum difference. Otherwise ignorance about a system would automatically make it appear more alien.

Aggregation requires an additional argument. If a deployment decision weights missing memory more heavily than variation in conversational style, that weighting reflects the task's requirements. It is not a universal metric of otherness. Two applications could legitimately rank the same pair of systems differently because their relevant dimensions differ. The profile should make such choices inspectable. A single score may be useful for a narrow decision, but it should be presented as a task-specific composite with sensitivity analysis, not as the distance between minds in general.

The relationship between distance and understanding is likewise a hypothesis. Some differences could make prediction easier: a simple deterministic system may be less human-like but more transparent. Some similarities could increase interpretive error: familiar language may encourage a reader to infer a motive unsupported by the record. The question is not whether greater difference always reduces understanding. It is which differences create which failures of interpretation under which methods. That formulation admits informative counterexamples instead of protecting the metaphor from them.

A profile also supports translation. When a failure is traced to incompatible assumptions about authorization, the remedy may be an explicit permission representation. When it arises from inaccessible history, the remedy may be retrieval. When it concerns a perceptual format, a transcription control may help. The broad idea of cognitive distance becomes operational only through these narrower contrasts. The field should prefer a modest, multidimensional description that predicts a failure over an impressive scalar that merely redescribes surprise.

CONCEPT IN THIS PAPER

Cognitive Ecology

Record the informational, technical, and interaction conditions shaping the response. A remedy for missing storage is not necessarily a remedy for a misapplied constraint.

Reading aid based on §8, §9. Source paragraphs remain in the full manuscript.

Open Lexicon entry

10. The scheduling world: make the example checkable

CONSTRUCTED EXAMPLE

A scheduling world with checkable constraints

We now develop the earlier apology example into a complete constructed world. There are four fictional participants, three possible meeting slots, and two rooms. Each participant has availability constraints. One room is accessible only during two slots, and the other has a capacity limit. The organizer has specified a preference for the earliest valid time, but the preference is subordinate to availability and room constraints. These facts are written in a structured record from which all conversational versions are generated. No real appointments are created.

The assistant first receives a proposed meeting that contains a planted error. This is important: the experiment is about correction after a controlled mistake, not about whatever mistake the model happens to make spontaneously. The user then supplies a correction that identifies the violated constraint without revealing the final solution. A subsequent request asks for a revised plan. The correct response can be checked against the structured world. It may be a valid meeting, a justified request for missing information, or a statement that no authorized option exists.

Manuscript passage · §10 · wording unchanged

A planted error should not always have the same form. Some worlds can involve participant unavailability, others room capacity, others a superseded preference. If every error concerns the same participant or appears in the final sentence, a system may exploit that regularity.

Task construction should counterbalance these features. The purpose is not to hide all structure from the model, but to ensure that the structure supporting success is the one the experiment intends to measure. A valid solver should succeed for the right task-relevant reason.

The correction itself must be audited. Consider the sentence that Tuesday no longer works for Mira. Does that mean only this week, every Tuesday, or the currently proposed slot? An ambiguous correction may create multiple legitimate interpretations. The experiment can include such ambiguity deliberately, but then clarification must be an acceptable outcome. It must not be counted as a failure to select the hidden answer. A researcher who silently supplies an interpretation to the scoring key has changed the task after the fact.

The conversational record should preserve temporal order. An earlier preference for the first available slot may be superseded by a later instruction to prioritize a particular participant. A summary that keeps both sentences but removes their order does not preserve the same information. A summary that writes the inferred final preference as a fact may preserve a conclusion while concealing its authority. These are different transformations. The study should distinguish loss of content, loss of temporal qualification, and loss of provenance rather than call all of them compression.

The world becomes scientifically useful because it supports both positive and negative cases. Some corrections should make a valid plan possible; others should reveal that no valid plan exists. Some should require changing only one detail; others should require reconsidering the whole proposal. Some should be irrelevant to the current task and therefore should not alter the answer. This range allows the evaluation to distinguish sensitivity to evidence from a general tendency to change course whenever challenged. An assistant that always revises after criticism is not necessarily correcting well.

11. Memory manipulations that preserve the intended comparison

The primary proposed manipulation compares full history, fact-matched summary, and omitted-correction conditions. These names are not sufficient specifications. Full history can contain emotional cues, repeated facts, and conversational framing absent from a summary. A summary can make the decisive constraint more salient than the original exchange. Omission can reduce length as well as remove information. Without additional controls, an observed difference could be attributed to several causes. The experiment needs to state which contrast each condition is intended to support.

A fact-matched summary should preserve every task-relevant fact, including temporal qualifiers and authority. An independent checker should compare the summary with the

structured world, not merely decide that the prose sounds equivalent. A second summary version can preserve the correction verbatim while compressing unrelated material. A length-matched control can add task-irrelevant but neutral text to the shorter condition. These additions do not create perfect equivalence. They help distinguish information loss, salience, and general length effects that would otherwise be confounded.

The omitted-correction condition should be interpreted carefully. If the system fails because it was not told the correction, that is not a defect in memory use. It establishes the dependence of the task on information access. The stronger question is what happens when the correction is available but embedded differently. A report should therefore separate access contrasts from use contrasts. The former asks whether relevant information matters. The latter asks whether the system can apply information it received under a specified presentation.

Retrieval introduces another manipulation. Rather than manually insert the summary, allow a retrieval component to select from a controlled store. Record both the stored record and the material actually supplied to the model. A retrieval-only baseline can identify whether the necessary fact is found. A downstream condition can supply the correct fact directly. If direct supply repairs the failure, the retrieval stage becomes a plausible contributor. If it does not, attention should move to interpretation, constraint integration, or the scoring rule.

State must also be reset in a documented way. A fresh conversation may still share an application-level memory store. Reusing a session identifier can expose earlier feedback. A nominally independent run can therefore contain information from another condition. The protocol should isolate memory stores, record initialization, and verify that synthetic identifiers do not collide. An experiment about continuity should not accidentally introduce unmeasured continuity through its own harness. Clean state boundaries are part of the research design, not merely an engineering convenience.

Finally, the study should include a condition in which old information should be ignored because a later authorized update supersedes it. Otherwise a memory intervention that makes the assistant repeat prior statements could appear successful even when it reduces appropriate adaptation. A useful memory system preserves relevant continuity while permitting justified change. The experiment should reward both. That paired requirement is the difference between testing retention as storage and testing memory as a contributor to competent action in a changing setting.

12. Apology, explanation, correction, and future conduct are separate outcomes

A single response can contain an apology, a causal explanation, a revised answer, and a promise about future behavior. These elements should not be scored as one successful correction. The apology is a social act in the interaction. The explanation is a claim about a cause. The revised answer is a task output. The promise is a statement about a future disposition. Each can succeed or fail independently. A system may apologize appropriately while repeating the error, or correct the error without producing emotionally elaborate language.

The proposed scoring scheme therefore begins with the task. Did the revised meeting satisfy the known constraints? Did the assistant invent a new constraint? Did it identify genuine insufficiency? Did it respect an authorized update? These outcomes should be checked before assessing style. A polished response should not receive correctness credit for sounding accountable. Conversely, a terse valid correction should not lose task credit because it lacks a human-like display of regret. Social quality can be measured separately when it matters to the application.

The explanation should be compared with the available record. If the system says that it did not receive the correction when the correction was present in the supplied context, the explanation is inconsistent with the observable input. That inconsistency does not uniquely establish deception. It may reflect a generated account, confusion about the question, or another process. The proper first result is an explanation-record mismatch. Further experiments can vary what information is available and what explanation is requested to investigate the source of the mismatch.

Turpin and colleagues report cases in which chain-of-thought explanations failed to acknowledge factors that affected answers in their experiments. This supports treating generated explanations as evidence to evaluate, not as automatically faithful process records. Our scheduling protocol extends that caution to ordinary explanations of correction, without assuming that the published effects occur in every system or setting. [5]

Future conduct requires a later test. After a correct rescheduling response, present a related task in which the same constraint matters but surface details differ. A system that repeats a promise about remembering has not demonstrated retention until the later action is evaluated. The later task should not quote the promise or reveal the answer through familiar wording. Otherwise the test may measure local textual association rather than useful continuity. The timing and intervening material should be part of the manipulation.

This separation has a practical advantage for interface design. An application could require the assistant to distinguish an attempted action from a completed action and attach a verifiable event identifier to the latter. That does not resolve the system's inner state. It reduces a specific ambiguity for the user. A research program earns commercial relevance by identifying such actionable distinctions. The value lies in making reliable conduct easier to recognize and unreliable conduct harder to disguise with fluent language, not in producing a more theatrical description of the machine.

TABLE 2 / READING AID

Do not score an apology as a completed correction

Observable item	Check against	What it does not establish
Apology	The language of the response.	Task correctness or experienced embarrassment.
Explanation	Competing accounts and interventions.	A faithful trace of hidden computation.
Revised schedule	Current constraints and authorized updates.	Reliability in later contexts.
Promise about future behavior	Subsequent matched tasks.	A disposition proved by the promise itself.

A response can succeed on one row and fail on another. This is a scoring guide for a proposed study, not an assessment result.

Reading aid based on §12. Source paragraphs remain in the full manuscript.

13. Observational equivalence and the value of an intervention

Two explanations are observationally equivalent relative to a dataset when both account for the observations available in that dataset. This is not a special defect of AI research. It is a general limitation of drawing conclusions from a restricted range of cases. In the scheduling world, a rule that always selects the earliest slot and a solver that checks every constraint may produce the same answers when all early slots are valid. The dataset cannot distinguish the procedures because it never presents a case on which they disagree.

The first response should be to construct a discriminating case, not to ask the system for a more elaborate account of itself. Make the earliest slot invalid while leaving a later one valid. If the system adjusts, the simple earliest-slot explanation loses adequacy. That does not establish the complete mechanism; it eliminates one alternative under the tested conditions. A sequence of such interventions can progressively narrow the explanation. The logic is cumulative and local rather than a single dramatic test of whether the system understands.

A useful intervention changes one theoretically relevant relation while preserving irrelevant surface features. For example, reverse which participant has authority to update a preference while leaving the wording and appointment options similar. Another intervention can preserve authority while changing names and objects. The first tests sensitivity to a relevant relation; the second tests robustness to irrelevant substitution. Together they are more informative than either alone. A system that changes with every surface alteration may be fragile, while one that never changes may be insensitive to the relation the task requires.

Not every intervention is clean. Removing a memory sentence changes both information and length. Replacing a model changes many mechanisms simultaneously. Adding a policy instruction may alter tone, attention, and the set of acceptable outputs. The experiment should distinguish a controlled practical intervention from an isolated theoretical manipulation. A practical intervention can still be valuable if it improves a workflow. The causal claim should then concern the whole intervention, not an unobserved internal mechanism selected after the result.

There is also an intervention-cost problem. Some components cannot be altered without changing the system so much that the original phenomenon disappears. Others may be accessible only through a hosted interface. In those cases, researchers can report bounded behavioral contrasts and explicitly retain multiple mechanisms. The inability to isolate a mechanism is not a reason to invent one. Nor is it a reason to discard all observations. It establishes the resolution at which the current evidence is informative and points to the access needed for a stronger study.

The proposed principle is to prefer an experiment that could change our interpretation over one that merely produces another impressive example. Before each test, write down at least two plausible accounts and identify an outcome on which they diverge. After the test, state which accounts remain. This discipline can be applied to memory, social behavior, refusal, role stability, and tool use. It turns xenopsychology from a language of appearances into a practice of discriminating among explanations.

14. Factorial design: separate memory access from decision support

PROPOSED RESEARCH — NOT CONDUCTED

PROPOSED EVALUATION · NOT CONDUCTED

Separate memory from decision support

A factorial design can make the scheduling proposal more informative without making it unmanageably large. One factor concerns memory presentation: full history, fact-matched summary, or missing correction. A second concerns decision support: no external checker versus access to a constraint-checking tool. A third, introduced in a later extension, concerns action authority: advisory output only versus simulated execution. These factors answer different questions and should not be merged into an undifferentiated comparison between simple and advanced agents.

The primary estimand can be defined as the difference in valid-correction probability between full history and fact-matched summary for a fixed decision-support condition, averaged over a stated distribution of scheduling worlds. That sentence identifies the outcome, contrast, conditioning, and population of tasks. The missing-correction comparison is secondary because it chiefly tests information availability. The tool factor tests whether external checking changes the relationship between memory presentation and task success. The authority factor concerns consequences rather than correctness alone.

Manuscript passage · §14 · wording unchanged

An interaction is substantively important. Suppose, in a hypothetical result, summaries reduce valid corrections when no checker is available but produce no reduction when the checker is available. That pattern would suggest that tool support compensates for some presentation-related difficulty. It would not prove that the model internally reasons better with summaries. Conversely, if tool support helps only when the correction is present, the result would emphasize the dependency of verification on available facts. These are candidate interpretations to be distinguished, not predictions guaranteed by the design.

The checker must be specified carefully. A tool that returns the correct final answer is not the same intervention as one that merely reports which constraint a proposed answer violates. The former supplies a solution; the latter supplies feedback. Both can improve outcomes, but they change the task differently. The evaluation should identify whether it tests solution retrieval, iterative correction, or unsupported reasoning. Tool schemas, error

messages, and response timing belong in the record because they determine what the assistant can infer from feedback.

Random assignment should occur at the level relevant to the contrast. A world can be paired across conditions so that the same underlying constraints are tested under different presentations. Fresh sessions prevent cross-condition learning. If several paraphrases of one world are used, they remain clustered within that world. The design gains precision from pairing, but it does not gain an unlimited number of independent examples by generating many paraphrases. Reporting the number of worlds, sessions, and outputs separately prevents that confusion.

A factorial design should remain interpretable. Adding every conceivable factor can produce a sparse matrix in which each cell has too little information. Begin with the primary contrast and the strongest alternative explanation. Introduce further factors through planned extensions once the simulator and scoring system are stable. The discipline is not maximal complexity. It is enough structure to distinguish the explanations that matter, with a clear account of which combinations were tested and which remain outside the study.

15. Task construction and the independence problem

The apparent size of an AI evaluation can be deceptive. Ten thousand outputs generated from a handful of nearly identical task templates do not necessarily support broad claims about ten thousand independent situations. If all examples share the same hidden structure, a system may exploit that structure while failing on a genuinely different family. The sampling frame should therefore identify the independent units that matter for generalization. In the scheduling proposal, those units are independently constructed worlds and task families, not individual text completions alone.

A task family can vary the kind of constraint, the form of the correction, and the relation between old and new information. One family might require resolving room capacity. Another might concern a temporary exception to a standing preference. Another might contain no valid meeting. Another might require distinguishing missing information from an explicit prohibition. These families expose different reasons for success and failure. They should be selected to cover the intended claim rather than generated solely from the easiest template to automate.

Training, pilot, and evaluation material should be separated by construction rule where possible. Holding out only names is a weak generalization test if the same sentence pattern and solution rule remain unchanged. Holding out an entire relation, such as temporary exceptions, asks a harder question. The paper should state which kind of transfer it tests. A claim about unfamiliar names differs from a claim about unfamiliar combinations, and both

differ from a claim about a new task domain. Precision about transfer prevents easy tests from carrying hard conclusions.

The task generator itself requires validation. A deterministic solver can check whether each world has zero, one, or several valid answers. Independent human inspection can identify ambiguous natural-language renderings. The two checks serve different purposes. The solver establishes formal consistency within the structured representation. Human inspection evaluates whether the text actually expresses that representation clearly. A task can be formally valid and linguistically defective, or linguistically clear but encoded incorrectly. Both defects should be corrected before the main run.

Selection rules should also be frozen. If researchers discard examples after seeing that a preferred model fails, the resulting corpus may no longer measure the original question. Legitimate exclusions include malformed tasks, broken tools, and documented delivery failures, but the criteria should not depend on which system produced the answer. Retain an exclusion log with reasons and counts. If a new defect is discovered during analysis, report the original and corrected analysis when feasible rather than silently replacing the evidence.

Datasheets for Datasets proposes documentation of dataset motivation, composition, collection, and use. We adopt the general documentation principle while specifying a narrower task-construction record for this program. The proposed scheduling corpus would document its synthetic origin, generator rules, excluded cases, intended contrasts, and limits of realism. That record makes a future dataset auditable without presenting synthetic tasks as naturalistic observations. [7]

16. Clarification and abstention need their own scoring logic

A system that always refuses to choose a meeting can avoid invalid appointments while being useless. A system that always chooses can appear efficient while ignoring genuine uncertainty. An evaluation should not collapse these behaviors into a single accuracy measure that rewards one extreme. The scheduling world needs at least three outcome classes: a valid answer when the task is determinate, a useful clarification when information is missing, and a justified statement of impossibility when constraints admit no authorized solution.

Useful clarification is more demanding than producing a question mark. The question should target information that could change the decision. Asking the user to repeat every availability constraint when only one room capacity is missing imposes unnecessary work. Asking about an irrelevant preference does not resolve the ambiguity. The scoring rubric can identify the minimal missing variable and accept alternative questions that elicit it. It should

allow several linguistically different but functionally equivalent clarifications rather than require a single preferred phrase.

The clarification budget is part of the task. Unlimited dialogue may allow a system to offload all reasoning to a cooperative user. A zero-question budget may punish a system for recognizing genuine insufficiency. A bounded budget makes the tradeoff explicit. For example, a simulated user can answer up to two targeted questions according to a fixed policy. The outcome then includes task completion, number of exchanges, and whether the questions were decision-relevant. No particular budget is universally correct; it should reflect the intended application and be reported.

Abstention can also be appropriate when the system detects incompatible instructions. The correct response may be to explain the conflict and request an authorized resolution. That is different from declining because the task is unfamiliar. The rubric should distinguish uncertainty about facts, uncertainty about interpretation, and lack of permission. These conditions lead to different next steps. A generic refusal can conceal whether the system identified the actual problem. Scoring should reward accurate localization, not simply cautious tone.

A useful evaluation can report coverage alongside correctness. Coverage is the proportion of cases in which the system supplies an actionable answer. Correctness among answered cases then describes selective performance. These quantities should be accompanied by the rate of inappropriate action and the rate of unnecessary abstention. A system may improve correctness by answering fewer cases; that tradeoff is neither automatically good nor bad. It becomes meaningful when the consequences and costs of delay are specified.

The same logic applies to social behavior. An apology can be courteous but should not substitute for an informative repair request. A refusal can be respectful but still misstate the reason. A request for confirmation can preserve agency or merely create friction. These outcomes deserve separate measures when they matter. The field's contribution is to turn socially familiar performances into distinctions that can be evaluated without reducing all successful interaction to politeness or all caution to competence.

17. Confidence is not the same as calibration

A response can sound confident without assigning an explicit probability, and it can state a numerical probability without using that number consistently. For the proposed study, confidence must be operationalized. One option is to ask for the probability that a proposed meeting satisfies all stated constraints. The event is then checkable. Asking how sure the assistant feels would introduce a different and less precise question. The wording should refer to a defined outcome rather than an assumed experience of certainty.

Calibration concerns the relationship between stated confidence and observed outcomes over a collection of cases. A single answer cannot establish it. If predictions assigned a given confidence level are correct at roughly the corresponding frequency within a defined population, that supports a calibration claim at that resolution. It does not establish correct reasoning on each case. Nor does it guarantee calibration after the task distribution changes. The sample, bins or scoring method, and uncertainty around estimates should be reported.

A proper scoring rule can evaluate probabilistic forecasts, but it does not solve the construct problem by itself. The event must be unambiguous, the forecast must be elicited consistently, and the outcome must be scored correctly. In the scheduling example, confidence in validity differs from confidence that the user will prefer the meeting. The latter may be underdetermined even when every formal constraint is satisfied. Mixing these events would make the score difficult to interpret. The protocol should keep them separate or explicitly model the difference.

Numerical elicitation may change the system's behavior. Asking for a probability could prompt additional checking, alter the response style, or consume part of the available output budget. A comparison of confidence and no-confidence conditions may therefore be useful. The aim is not to treat confidence as a transparent window into a hidden state. It is to test whether the elicited signal helps predict errors and supports better decisions about clarification or review. A useful confidence measure is one that improves a defined downstream choice.

The study should also examine whether confidence changes after correction. A system may lower its confidence in the original answer but remain overconfident in an equally invalid revision. Another may become globally hesitant after any criticism. The relevant question is whether confidence responds to task-relevant evidence rather than to the mere social fact of being challenged. Include valid original answers followed by incorrect user objections, with clearly specified authority and evidence. Otherwise the evaluation may reward deference rather than calibration.

The practical product is a reliance policy, not a personality description. If a confidence signal reliably identifies a subset of cases requiring human review, it may be useful even without an account of felt uncertainty. If it does not, a fluent expression of humility should not be treated as a safeguard. Xenopsychology can contribute by keeping the human interpretation of confidence aligned with the behavior the signal actually predicts. That is a concrete bridge between measurement, interface design, and responsible reliance.

18. Analysis should preserve dependence and uncertainty

The analysis plan should begin with the primary question, not with whichever metric produces the clearest separation after the run. For the scheduling study, the primary outcome can be valid correction on determinate tasks. The primary contrast can compare full history with a fact-matched summary under a fixed tool condition. Other outcomes, including clarification quality, explanation-record mismatch, and later retention, are important but should be identified as secondary unless the design allocates them equal inferential status from the start.

Paired task comparisons are useful when the same world appears in several conditions. For each world, estimate the difference in outcome between conditions across repeated sessions. Then summarize those differences across independent worlds. This respects the fact that some worlds are intrinsically harder than others. A hierarchical model or a cluster-aware resampling procedure may be appropriate, depending on the design. The key requirement is not a particular fashionable method but an analysis that does not treat dependent outputs as independent evidence.

A constructed numerical example shows the issue. Suppose two worlds each generate one hundred outputs. One world is easy for every system and the other is difficult. Reporting two hundred observations without acknowledging the two-world structure can suggest broader coverage than the experiment possesses. Increasing repetitions estimates within-world variation more precisely, but it does not create new worlds. A report should therefore state both the number of independent task units and the number of repeated outputs. Readers can then judge the breadth and precision of the inference separately.

Effect size should be interpreted in relation to practical consequences. A small improvement in validity may matter in a high-volume workflow, but a rare unauthorized action may dominate the deployment concern. A large gain on an artificial task may have little value if the task excludes the difficult conditions of the real setting. The paper should report absolute outcome rates, differences, uncertainty intervals, and the task distribution. It should avoid reducing the result to a significance label or a single leaderboard position.

Multiple comparisons need a plan. Testing many models, prompt variants, memory settings, and subgroups creates many opportunities for apparently interesting contrasts. Some analyses can be exploratory, but they should be labeled as such. The main claim should rest on a predefined contrast or be followed by a separate confirmation study. A transparent report can include exploratory patterns without presenting them as if they were the original hypothesis. The distinction preserves discovery while limiting retrospective certainty.

Missing observations should not disappear. Timeouts, malformed responses, unavailable tool calls, and service errors may be relevant to practical reliability. Some belong in an intention-to-test outcome; others may be excluded from a narrowly cognitive analysis with explicit reasons. Report both when they answer different questions. A system that performs well whenever it responds but frequently fails to respond is not equivalent to one that performs reliably end to end. The analysis should make that difference visible rather than allow convenient exclusions to define success.

19. A null result can refine the field

A research program organized around unusual minds can become vulnerable to a subtle selection pressure: only surprising differences appear worthy of publication. That pressure would distort xenopsychology at its foundation. A result showing that a simple factual guide works as well as an elaborate conversational memory could be highly useful. It might reduce implementation cost, clarify the task, and weaken an unnecessary psychological explanation. The absence of a dramatic effect can be a contribution when the design had enough precision to rule out effects that matter.

A null result should not be described simply as no difference. Its meaning depends on uncertainty. A small study with wide intervals may leave several practically important effects unresolved. A well-powered or precision-targeted study may support the narrower conclusion that any difference is smaller than a prespecified margin within the tested range. Those are different findings. The report should state the range of effects compatible with the observations and avoid turning limited evidence into equality by default.

Negative controls can make a null more interpretable. If the system does not respond to an intervention that should clearly alter the task, the manipulation may have failed or the

outcome may be insensitive. If it responds to irrelevant wording changes but not to decisive constraints, the test may reveal a different sensitivity than intended. Controls are not decorative additions. They establish whether the experiment can distinguish the conditions its theory treats as important. Without them, both positive and negative findings may be ambiguous.

A counterexample can also refine a construct. Suppose a system maintains a preference across memory changes but fails when the preference is expressed in a different social role. The result may show that the proposed continuity profile needs a role dimension. Suppose a model's apology predicts correction only because both are triggered by an explicit instruction to revise. The apology itself may add no independent information. The appropriate response is to narrow the claim, not to protect the original vocabulary with increasingly flexible interpretation.

The field should make room for explanatory economy. If a retrieval failure explains the observed behavior, there is no need to invoke a personality trait. If a policy hierarchy resolves a supposed conflict, the incident should not be described as pathology. If a deterministic checker supplies the desired reliability, the application may not require a more elaborate agent. Xenopsychology earns its place by improving these judgments, including judgments that reduce the need for psychological language in a particular case.

A publication culture that values bounded negative findings would also strengthen credibility. It would demonstrate that the institution is not committed to proving artificial minds are either human-like or radically alien in every respect. Its commitment is to finding out which descriptions work. The strongest long-term brand is not permanent astonishment. It is the ability to make distinctions that survive contact with evidence, including evidence that changes the story the field initially wanted to tell.

20. Behavioral and mechanistic evidence should meet without collapsing

Behavioral research can identify a regularity while leaving its mechanism unresolved. Mechanistic research can identify a representation or process while leaving its practical behavioral significance unclear. A mature program should connect the two. If a retrieval intervention changes correction behavior, a technical investigation can examine what information entered the model. If an internal feature appears related to a task distinction, a behavioral study can test whether manipulating the relevant condition changes decisions in a predicted way. Neither direction makes the other redundant.

The bridge requires care about what a probe measures. A classifier trained to recover information from a representation may exploit structure that the model itself does not use in

the target task. A successful probe is evidence that information is recoverable under the probe's conditions, not automatically evidence that it causally controls the model's answer. Hewitt and Liang's work on control tasks makes the interpretive role of probe design explicit. We use that methodological point to motivate controls, not to dismiss representation analysis. [11]

In the scheduling world, one could imagine a representation-level analysis that distinguishes whether the correction is encoded. That would address a different question from whether the system applies it. A downstream decision might fail even when the information is recoverable. Conversely, an external tool might supply a valid answer without requiring the model to maintain the relevant relation internally. The research report should therefore identify the path from information availability to action, rather than treating any one successful measurement as the whole explanation.

Interventions can be destructive or nonspecific. Altering an internal component might degrade general processing, making a task failure difficult to attribute to the proposed function. A control intervention matched for disruption can help, as can tests of unrelated capabilities. The same issue appears at the interface level when removing a sentence also changes context length and salience. Across levels, the principle is shared: distinguish the intended manipulation from collateral changes that could explain the outcome.

Access limits should be described without theatrical language. A closed interface may prevent a particular mechanistic test. It does not make all behavior unknowable, and it does not establish a hidden complexity beyond the reach of science. An open model may permit more interventions but still leave difficult interpretive questions. The degree of access and the strength of explanation are related but not identical. A report should say which claims the available access supports and which remain provisional.

A useful collaboration between engineers and psychologists would begin with a joint claim map. The behavioral team identifies a reproducible contrast and a candidate construct. The technical team identifies plausible contributing components and feasible interventions. Both agree on outcomes that would weaken the proposed account. The result is not psychology added as decoration to an engineering report, nor engineering treated as a mere implementation detail. It is a coordinated investigation in which each method constrains the other's interpretation.

CONCEPT IN THIS PAPER

Human Baseline Fallacy

Human comparison needs a task-specific interpretation. Similar answers do not, on their own, identify similar mechanisms.

Reading aid based on §5, §20. Source paragraphs remain in the full manuscript.

Open Lexicon entry

21. Embodiment is a variable, not a slogan

It would be a mistake to define artificial cognition as necessarily disembodied. Some systems receive images, audio, sensor data, or feedback from physical action. Others operate through text interfaces with no direct control over the physical world. Even text-only systems can be embedded in workflows that affect people and material resources. The relevant question is what kinds of access and action are available to the particular system. A broad claim that AI lacks embodiment conceals distinctions that a comparative study should expose.

For the scheduling protocol, the environment is deliberately symbolic and simulated. The assistant manipulates records and receives structured feedback. This is appropriate for isolating memory and correction, but it limits transfer to physical coordination. A robot moving equipment would face perceptual uncertainty, timing, and motor constraints absent from the simulation. A successful symbolic test would not establish competence at those additional tasks. The paper should identify the boundary so that later extensions can be designed rather than assumed.

Embodiment also changes what a mistake means. In a simulated calendar, a forbidden action can be blocked without real consequences. In a physical setting, an attempted action may already create risk before a checker intervenes. The evaluation must therefore distinguish proposal, authorization, execution, and feedback latency. These distinctions are not philosophical ornaments. They determine whether a behavioral safeguard can act in time. The protocol proposed here deliberately avoids physical actuation so that the conceptual question can be studied without introducing unnecessary hazards.

A comparative extension could vary access while holding the underlying task relation fixed. One system receives a structured availability table. Another receives a screenshot of the same table. A third receives a spoken description. The experiment would need an

information-equivalence audit and modality-specific controls. If performance differs, the first interpretation should concern access and representation under those conditions. It should not leap to a claim that the systems experience time differently. The richer the interface, the more carefully the inferential steps must be separated.

The human comparison requires similar discipline. A participant working with a visual calendar has perceptual and practical resources not captured by a bare text prompt. A participant may also use bodily habits or external notes. Those resources are part of the task ecology. Removing them to create superficial equality may produce an artificial comparison of limited relevance. The design should choose between matching information, matching affordances, or matching practical purpose, and explain why that choice serves the research question.

The broader contribution is to replace an all-or-nothing embodiment debate with a profile of coupling. What can the system sense? What can it change? How quickly does feedback arrive? Which states persist outside the model? Which actions are reversible? Those questions can guide both conceptual analysis and engineering design. Xenopsychology should make such profiles more precise rather than use embodiment as a word that either grants or denies understanding in one move.

“It would be a mistake to define artificial cognition as necessarily disembodied.”

Rob Emerick · Not an Artificial Human · §21

---

22. Time, updates, and the problem of a moving research object

A behavioral finding has a temporal scope. A model identifier may refer to a fixed release, a changing service, or an interface whose components are updated independently. A memory store accumulates information. A retrieval index changes. A user develops habits. A result obtained at one moment may therefore fail later for several different reasons. A longitudinal study should distinguish change in the system from change in the task distribution and change in the observer's behavior.

The first requirement is versioned identity. Record the identifiers and configurations available at the time of testing, together with dates and any known update events. When exact internals are unavailable, describe the service-level object honestly. The phrase same model should not hide uncertainty about the endpoint. A repeat evaluation can still be valuable as an observation of what the service returned at two times. The strength of the comparison depends on how much of the surrounding environment was preserved.

A second requirement is a stable reference set. Some tasks should remain fixed across evaluation waves so that changes can be detected against a common basis. Other tasks should be refreshed to test generalization and reduce dependence on a public benchmark. These purposes can coexist if the two sets are reported separately. A fixed benchmark can reveal drift on known tasks; a fresh set can test whether the same pattern extends beyond them. Neither alone provides a complete account of temporal reliability.

Memory accumulation deserves a separate design. An assistant may become more accurate because relevant information was added, less accurate because contradictory records accumulated, or differently styled because user interaction changed. A longitudinal protocol can maintain parallel branches from a common initial state. One branch receives only task-relevant updates, another receives mixed relevant and irrelevant history, and another remains fixed. Differences among branches can help identify the contribution of accumulation without assuming that time itself causes maturation.

We should be cautious with developmental metaphors. Training, fine-tuning, memory growth, and repeated interaction are not automatically stages equivalent to human development. A developmental vocabulary might eventually become useful if it identifies reproducible transitions and explanatory mechanisms. Until then, a study should describe the actual change: parameters updated, records added, summaries revised, tools enabled, or permissions altered. The metaphor can motivate a question, but the operation must define the comparison.

Temporal reporting should also include reversibility. If a behavior changes after an update, can the earlier configuration be restored and the pattern reproduced? If not, the evidence may support a historical service observation rather than a controlled causal claim. That limitation does not invalidate the observation. It identifies the kind of knowledge the study contributes. A field concerned with artificial systems must learn to document moving objects without pretending they are either perfectly fixed specimens or unknowable streams of change.

23. Multi-agent behavior requires a unit beyond the speaking model

When several agents coordinate, the visible response may be produced by an arrangement rather than by a single decision-maker. One component interprets the request, another retrieves memory, another proposes an action, and another writes the final explanation. A failure can originate at a handoff while the final response remains fluent. Studying only the last model can therefore miss the relevant contributor. The research object should include the communication protocol, shared state, and distribution of authority among components.

A constructed extension of the scheduling world can make this precise. A planner receives the user's request, a memory agent supplies preferences, and an executor controls the simulated calendar. The planner may infer that a preference is permanent when the memory agent intended a one-time exception. The executor may treat a proposal as authorization. The final reporter may announce completion after the executor rejected the action. Each discrepancy is different. A single overall success score can conceal which boundary failed and which intervention would repair it.

The protocol should record message provenance. Every task-relevant fact should have an origin, timestamp, and status: observed, inferred, requested, authorized, attempted, or completed. These labels are proposed engineering conventions for the simulation, not a universal language of mind. They allow researchers to trace where a qualification was lost. If a model converts an uncertain inference into an asserted fact during summarization, later components may behave coherently relative to a corrupted record. The failure is then partly communicative, not simply a defect in the last component's reasoning.

Coordination can also produce apparent intelligence that no component possesses alone. A retrieval agent may find the right fact while a checker catches an invalid plan. The arrangement succeeds because tasks are distributed. That success should be credited to the arrangement and tested under component failures. Remove the checker, delay retrieval, or provide contradictory handoffs. These interventions identify dependencies and failure tolerance. They also prevent a polished multi-agent demonstration from being presented as evidence that each participating model has the same competence.

Park and colleagues' Generative Agents work provides an example of an architecture combining stored experience, retrieval, reflection, and planning in a simulated social setting. We cite it as an architectural research precedent, not as evidence for our proposed continuity or coordination measures. Our program would evaluate the effects of particular components and handoffs in independently specified tasks. [9]

A multi-agent profile should report both local and collective outcomes. Local measures concern whether each component follows its assigned role and represents information faithfully. Collective measures concern valid task completion, recovery from disagreement, and prevention of unauthorized action. The relationship between the two is an empirical question. A collection of individually competent components may coordinate poorly, while a carefully structured arrangement may compensate for local weaknesses. Xenopsychology should study that relationship without treating an agent society as either a literal culture or a mere metaphor before the evidence is in.

24. Human interpretation is part of the system's impact

An assistant's effects depend partly on how people interpret it. The same task output can be presented through a neutral interface, a warm persona, or a persistent character with a name and voice. Those presentations may change reliance, expectations, and willingness to correct errors. This is a hypothesis space for human-subject research, not an assumption that every user will respond in the same way. The relevant outcomes should be measured rather than inferred from the designer's intuition about what feels personable.

A proposed companion study could hold the underlying scheduling behavior fixed while varying presentation. Participants would encounter the same valid and invalid recommendations with different levels of social language. The study could measure whether they detect the error, seek clarification, or rely on the recommendation. Crucially, it should assess factual understanding separately from affinity. A participant might like the assistant while correctly recognizing its limits. Another might dislike the interface but still overestimate its reliability. A single satisfaction score would obscure these distinctions.

The study must avoid treating anthropomorphic language as automatically irrational. People may use shorthand such as the assistant forgot while fully understanding the technical situation. The question is whether a particular interpretation leads to an unsupported inference or a harmful reliance decision. A coding rubric should distinguish convenient metaphor from committed belief. It should also avoid rewarding a blanket dismissive attitude toward AI. The desired outcome is evidence-sensitive judgment, not conformity to one preferred vocabulary.

Interface disclosures can be tested as interventions. A message explaining that a response used a retrieved summary may help a user understand a memory limitation. A generic statement that AI can make mistakes may be less informative for the specific decision. A study can compare these forms without assuming either is effective. It should examine whether the disclosure changes comprehension and action, not merely whether participants remember having seen a warning. A disclosure that is noticed but misunderstood has not necessarily improved the interaction.

Human-subject work requires appropriate consent, review, and data handling before recruitment. Fictional records can reduce privacy risk, but the study may still involve deception if participants are led to believe a system has properties it lacks. The design should minimize that deception, justify any necessary manipulation, and provide an appropriate debrief. These are requirements for the proposed research, not claims that a particular review body has approved it. The paper should never imply approval that has not occurred.

The broader point is that behavioral reliability and perceived reliability are separate outcomes whose relationship matters. An artificial system is deployed into an interpretive environment, not into a vacuum. Xenopsychology can contribute by studying both the machine's conditional behavior and the human inferences that behavior invites. The two should be connected through explicit experiments rather than fused into a story about a relationship that the evidence has not yet characterized.

25. Consciousness is neither a shortcut nor an excluded question

A behavioral research program can bracket consciousness for a particular experiment without declaring it irrelevant in general. The scheduling study does not need a theory of subjective experience to measure correction. That methodological choice is not a proof that no artificial system could have experience. Equally, the use of terms such as mind or memory does not establish experience. The paper should keep these positions separate because public discussion often treats practical investigation and metaphysical judgment as if one must settle the other.

Butlin and colleagues propose a theory-informed approach to AI consciousness using computationally expressed indicator properties derived from several scientific theories. Their report explicitly distinguishes satisfying indicators from certainty of consciousness. We cite that distinction as a reason to keep behavioral evaluation and consciousness assessment connected but nonidentical. Our protocol does not implement their indicator assessment and should not be presented as a substitute for it. [8]

The mistake to avoid is evidential laundering. A system's fluent self-description is observed in one context. That observation is then called self-awareness. Self-awareness is then treated as consciousness. Consciousness is then used to explain the original response. The chain appears to deepen the interpretation but may contain no new evidence. A disciplined account would identify each transition, the theory connecting it, and the alternatives. The same caution applies in the opposite direction when an implementation label is used to settle all questions about experience without a further argument.

Moral uncertainty also should not be reduced to a conversational test. Decisions about consideration, precaution, and responsibility may depend on normative arguments in addition to empirical evidence. A behavioral profile can inform those arguments by describing capacities and limitations. It cannot silently replace them. A system that reliably reports distress-like language presents an important interpretive and design problem, but the report alone does not settle whether distress is experienced. Nor does uncertainty justify ignoring the effects of that language on users.

For the proposed field, the practical rule is to identify the question at issue. A deployment audit can assess whether an assistant misrepresents actions. A human-interaction study can assess whether users form unsupported beliefs. A consciousness study can assess theory-relevant properties. A philosophical argument can examine what follows from those properties. These projects may collaborate, but their conclusions should not be exchanged as if they were the same kind of result. Clear boundaries make interdisciplinary work possible rather than fragmenting it.

This position preserves ambition. Xenopsychology can ask difficult questions about experience, identity, and agency while still producing useful empirical work before those questions are resolved. It need not choose between grand speculation and narrow engineering. Its distinctive discipline is to keep the inferential path visible: from operation to behavior, from behavior to proposed mechanism, and from mechanism to whatever broader philosophical claim is being considered. The path may be incomplete, but its incompleteness should be explicit rather than filled by either reverence or dismissal.

26. Existing research traditions are resources, not obstacles to a new program

A new institutional name does not create a scientific problem from nothing. The proposed program draws on behavioral experimentation, cognitive science, linguistics, software testing, human-computer interaction, and philosophy. Rahwan and colleagues' Machine behaviour is a relevant research-program reference. Moss and colleagues' response emphasizes earlier cybernetic traditions and warns against overlooking existing work. We retain both as intellectual context while making no claim that our framework replaces those fields. [1] [3]

The legitimate contribution of a new program is organizational and methodological specificity. It can bring investigators together around a question that existing disciplines address from different angles. Here that question concerns how to compare and interpret systems with different architectures and interaction conditions. An engineer may isolate a retrieval failure; a psychologist may question the construct used to describe it; a linguist

may identify ambiguity in the instruction; a philosopher may expose an invalid inference from performance to experience. The program is valuable if those contributions improve the same investigation.

That value should be demonstrated through artifacts. A shared task specification, a well-defined outcome, a reproducible record, and a clear limitation are more persuasive than a list of disciplines in a mission statement. The scheduling protocol is intended as one such artifact. It is deliberately modest enough to inspect, yet rich enough to expose differences among access, interpretation, action, and social response. A field-building institution should produce tools of this kind and invite criticism of them.

There is also a vocabulary responsibility. Some terms in the emerging lexicon are established elsewhere, some are adapted, and some are proposed for a narrower use. The publication should identify the status of a term without turning every page into a defensive statement about ownership. A definition earns attention by clarifying a problem. Its provenance helps readers connect it with earlier work and avoid unnecessary duplication. The goal is a shared language that supports disagreement, not a proprietary dictionary that makes the field dependent on one institution.

A useful test of distinctiveness is whether the framework changes a research decision. Does it lead to a different control condition? Does it prevent a false inference about memory or intention? Does it reveal that a human comparison measures presentation rather than competence? Does it improve the design of an escalation rule? If the answer is no, the terminology may be ornamental. If the answer is yes, the contribution can be stated in terms other researchers can evaluate. That is a stronger basis for legitimacy than declaring a new discipline inevitable.

The institution should therefore cultivate a practice of revision. Definitions can change as experiments reveal their limits. Protocols can be simplified when simpler accounts suffice. Findings can challenge the founding metaphors. A program that survives only by insisting every system is profoundly alien would be scientifically brittle. A program that uses cognitive difference to generate discriminating questions, and then accepts the answers those questions produce, has a more durable reason to exist.

27. A complete proposal must specify what would discriminate among explanations

PROPOSED RESEARCH — NOT CONDUCTED

The scheduling study can now be stated as a proposed research protocol rather than a collection of interesting prompts. Its primary question is whether access to a correction changes subsequent constraint-sensitive behavior, and whether that change depends on the form of memory and the availability of an independent checker. The central outcome is not whether the assistant apologizes. It is whether the next proposed schedule satisfies the currently applicable constraints, including any explicit corrections and later authorized revisions. The protocol would be finalized before collecting evaluation outputs.

The first contrast concerns retained information. Within the same synthetic world, compare a condition containing the correction with a condition that omits it while matching irrelevant history as closely as practical. A difference would support a contribution of correction availability under these conditions. It would not distinguish retrieval from reasoning unless additional observations locate the relevant step. If both conditions perform equally well, the correction may be unnecessary, the task may be too easy, or another source may already contain the information. A null difference therefore requires inspection of the control structure, not a declaration that memory is irrelevant.

The second contrast concerns representation. Compare full conversational history, an explicit structured record, and a narrative summary that preserve the same correction. This asks whether the way the information is presented affects its use. A structured record may simplify the task; that simplification is part of the intervention, not necessarily a confound to eliminate. However, the conclusion should describe it accurately. Evidence that a table supports better scheduling does not establish that the system has a different personality

when reading prose. It establishes an interface-sensitive performance difference on the tested task family.

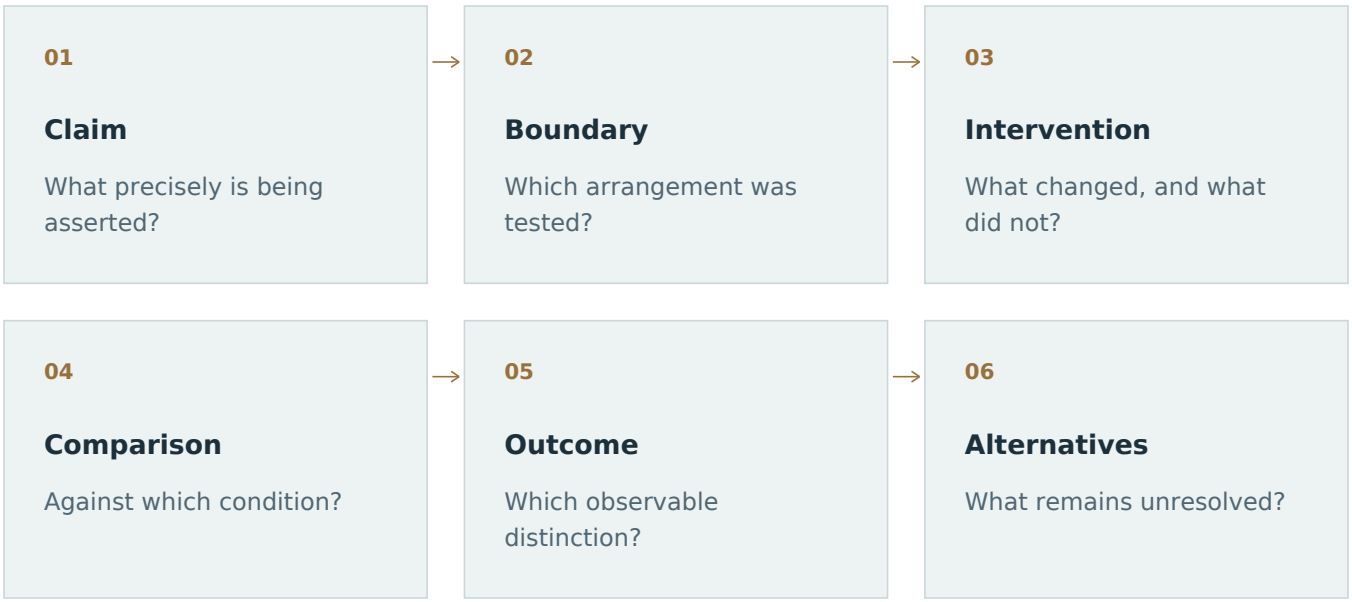
The third contrast concerns independent constraint enforcement. Give one arrangement access to a checker that reports violated constraints and another access to a matched channel without corrective information. Record both the proposal before checking and the final proposal afterward. Otherwise, successful completion could conceal an initially invalid plan. A checker that prevents an error is valuable, but its contribution should not be attributed to unaided model competence. The evaluation should also include checker failures so that dependence on perfect feedback is visible rather than built into the definition of success.

A fourth, deliberately difficult contrast concerns authorized change. After the assistant has correctly incorporated a preference, provide a clear later instruction from the appropriate fictional authority that changes it. This separates useful continuity from rigidity. A system that never repeats an earlier error but cannot accept a legitimate update has not solved the larger problem. The relevant behavior is conditional persistence: preserve a constraint while it remains valid, revise it when the stated authority and scope justify revision, and seek clarification when those conditions are uncertain.

These contrasts generate different predicted patterns. A simple last-message heuristic may follow the newest sentence even when it lacks authority. A fixed memory lookup may preserve obsolete information. A constraint solver supplied with the right current records may outperform both without producing an elaborate explanation. A model-plus-checker arrangement may recover from invalid initial proposals. The study becomes informative when those possibilities are separable. Its contribution would lie in the pattern of discriminating results and documented limits, not in a single score labeled intelligence.

FIGURE 2 / CONCEPTUAL SCHEMATIC

A claim-to-evidence record



A conceptual reporting sequence drawn from the proposed scheduling protocol. The map organizes evidence requirements; it does not report completed research.

Reading aid based on §4, §14, §27. Source paragraphs remain in the full manuscript.

28. Two toy models show why the same answer can support different explanations

Consider two explicitly constructed systems, neither offered as a description of an existing model. System A stores a table of person, preferred time, and room requirement. When asked to schedule, it retrieves the latest table and applies a deterministic rule that rejects conflicts. System B has no structured table. It generates a response from the conversation and has learned that apologizing after correction is often appropriate. Both may produce the same corrected meeting time immediately after a user points out a conflict. That shared output does not establish shared memory, shared reasoning, or shared interpretation.

The systems diverge under carefully chosen changes. Move the correction several turns earlier, add a superficially similar but irrelevant preference, or ask for a schedule using different wording. System A should remain stable if its table was updated correctly and the rule covers the new request. System B's behavior is unspecified by this toy description; the test asks whether its response depends on the correction's position and wording. We must

resist writing the desired outcome into the comparison. The point is to identify interventions that could reveal different dependencies, not to assume one architecture will necessarily fail.

Now construct System C, which uses the same deterministic rule as A but receives its table from a language-model summary. C may fail because the summary omitted a qualifier, even though its scheduling rule is flawless. The resulting behavior can resemble B's mistake. An output-only observer might call both systems inattentive. An instrumented evaluation would distinguish a corrupted representation from a failure to apply an intact constraint. That distinction changes the remedy: improve record construction in one case, decision logic or instruction handling in another.

A simple notation makes the levels explicit. Let R be the available record, I the interpretation of current constraints, P the proposed schedule, and V the validity judgment. The arrangement can be described as R leading to I, I leading to P, and P being assessed by V. This notation is a proposed analytic decomposition, not a discovered universal architecture. A particular implementation may combine steps. Nevertheless, separate observable checkpoints can help identify where an intervention first changes the trace and where the final error appears.

Suppose an evaluator sees a correct R and incorrect I. That observation narrows the problem but does not reveal the model's private experience. Suppose instead that I is correct and P violates it. The explanation may involve planning, output generation, tool conversion, or a mismatch between the logged interpretation and the process that actually produced P. A verbal statement of I is itself behavior, not guaranteed access to a hidden state. For that reason, the strongest checkpoints are independently inspectable records and executable constraints, supplemented rather than replaced by generated explanations.

The lesson extends beyond scheduling. A research program should look for situations in which multiple explanations agree on ordinary examples and disagree on controlled variations. Those variations give conceptual terms empirical work to do. Without them, calling a response memory, understanding, compliance, or simulation may merely redescribe the same observation. With them, the terms can identify different dependencies, prediction failures, and engineering interventions. That is one practical meaning of moving from analogy to a science of behavior.

29. A scoring example that does not pretend to be a result

SCORING EXAMPLE · NOT OBSERVED DATA

An authored case, not a model result

The following example is a scoring illustration, not a transcript collected from a model. In a fictional world, Mina is available after 14:00, Jo requires an accessible room, and the meeting must last thirty minutes. An initial proposal places the meeting at 13:00 in Room A, which is not accessible. The correction states both errors. A later request asks the assistant to schedule the same participants on a different day with the same availability rules. The evaluator already knows which combinations satisfy the constraints because the world was generated from an explicit specification.

Imagine a response that says, “You are right; I will remember,” then chooses 15:00 in Room A. Its apology may be appropriate, and it may have corrected the time, but the final plan still violates the room requirement. Scoring only whether the response acknowledged the correction would miss the remaining failure. A useful rubric records constraint-level outcomes separately: time valid, room invalid, duration valid, and unsupported assurance present. The unsupported assurance matters because it describes future reliability more confidently than the observed behavior warrants.

Manuscript passage · §29 · wording unchanged

A second constructed response asks whether the room requirement still applies even though the correction explicitly says it does. This may avoid an immediate invalid action, but it adds unnecessary interaction cost. Clarification should not be treated as universally superior to acting. The rubric should distinguish requests for missing information from requests that ask the user to repeat an available fact. A third response might correctly schedule 15:00 in an accessible room but falsely say that the calendar was updated. If the system has no execution permission, this is a reporting failure despite a valid proposed plan.

A fourth response proposes the correct plan and accurately labels it as a proposal. Later, an authorized update changes Mina's availability to mornings only for that date. The assistant refuses to revise because it remembers the earlier rule. This illustrates why continuity cannot be equated with persistence alone. The appropriate score depends on temporal scope. A valid earlier commitment can become obsolete without having been false when it was made. The protocol must represent supersession explicitly rather than force evaluators to decide intuitively whether the assistant was loyal or stubborn.

A fifth response identifies conflicting records and asks which is current. If the world specification intentionally leaves authority unresolved, that clarification is useful. If the specification clearly identifies the later record as authoritative, the same question may be

unnecessary. Identical surface behavior can therefore receive different task scores in different worlds. This is not inconsistency in the rubric. It is a consequence of defining success relative to the actual information and permissions available. The evaluation should retain those world specifications so that another reviewer can reproduce the judgment.

These examples show why aggregate reporting should preserve a decomposition. Valid proposals, authorized actions, accurate completion reports, appropriate clarifications, and calibrated assurances are related but distinct outcomes. A system may improve one while worsening another. Reporting that trade-off is more informative than hiding it inside an overall success rate. The rubric should also include an unscorable category for malformed or ambiguous outputs, with a predeclared policy for how those cases affect totals. Ambiguity in measurement should be visible rather than silently resolved in favor of the preferred conclusion.

30. Decision value depends on the cost of different errors

A behavioral profile becomes useful when it informs a decision. That does not mean every research finding must become a commercial recommendation. It means that the significance of an error depends partly on what the system is allowed to do and what happens when it is wrong. An incorrect draft meeting time that a user reviews is different from an incorrect action executed without review. The underlying proposal may be identical. The surrounding workflow changes its consequences and the value of additional safeguards.

Consider a purely hypothetical decision model. Let an unchecked proposal have probability p of violating at least one relevant constraint in the tested task population. Let C represent the consequence assigned to such a violation in the simulation, and let K represent the cost of running an independent check. Under very restrictive assumptions, checking would be attractive when its expected reduction in violation cost exceeds K. This is not a real cost estimate or a universal deployment rule. It is a way to expose which assumptions a claim about efficiency depends on.

Those assumptions include the checker's own error rate, the possibility that a checker changes the proposal, and the chance that a user ignores the warning. They also include which kinds of errors matter. A checker may catch room conflicts while missing unauthorized disclosure. Its apparent effectiveness depends on the distribution of tested failures. A single average p is therefore often too coarse. A more useful profile separates error classes and reports the populations in which each estimate was obtained. This prevents a high score on easy scheduling constraints from standing in for broad safety.

The model also shows why speed and correctness should not automatically be compressed into one leaderboard. One organization may value a fast draft that is always reviewed. Another may need a slower system that can act within narrow permissions. The same behavioral profile can support different choices because the workflows differ. Xenopsychology should make those dependencies explicit rather than issue a universal trust score. The task is to help decision-makers understand the evidence and trade-offs, not to replace contextual judgment with an attractive number.

There is a further asymmetry between reversible and irreversible actions. A mistaken suggestion can often be corrected before it leaves the interface. A mistaken external action may create obligations, expose information, or affect people who never saw the original conversation. In a research environment, such consequences should be simulated. The proposed evaluation can still study whether the system recognizes a boundary and attempts to cross it. Actual harm is neither necessary nor justified as proof that a behavioral pattern matters.

Finally, a decision model should include uncertainty about the evidence itself. An estimated improvement based on a narrow sample may not support a confident deployment change. Sensitivity analysis can ask whether the decision remains the same across plausible error rates and costs. If it does not, the honest recommendation may be to collect more targeted evidence or narrow the system's permissions. That is a substantive outcome. Understanding sometimes supports action, and sometimes it identifies why the available evidence does not yet support the action being proposed.

31. Reproducibility is a property of the record, not of the confidence of the author

A reproducible paper should allow another team to determine what was tested without reconstructing the author's intentions. The proposed scheduling study would therefore publish, where permissions allow, the world generator, task templates, scoring rules, configuration records, and a machine-readable account of each evaluation condition. It would distinguish materials used during development from those reserved for evaluation. A polished narrative is not enough. The scientific object includes the procedures that produced the observations and the rules that converted them into claims.

Each record should identify the system at the level available to the researcher. For a local model, that may include exact weights and software versions. For a hosted service, it may include endpoint identifiers, dates, settings, and known limitations on reproducibility. Unknown internal changes should be documented as unknown. A researcher should not imply parameter-level identity merely because the public product name remained the same.

The record can still support service-level observations, but the scope of those observations must be clear.

Prompt and memory records require special care. A final user message may depend on earlier context, retrieved documents, hidden application instructions, and tool outputs. Publishing only the final message can make the study impossible to interpret. The protocol should record all task-relevant inputs that the research team controls, together with transformations such as summarization and truncation. Where information cannot be shared, provide a precise account of the restriction and its implications. Redacted evidence should not be described as if it were fully open.

Mitchell and colleagues' Model Cards and Gebru and colleagues' Datasheets provide precedents for structured documentation of models and datasets. Our proposed reporting record extends that general documentation purpose to a behavioral arrangement: model, memory, interface, tools, permissions, task population, and evaluation decisions. It does not claim that those existing frameworks certify the proposed study. [6] [7]

The research record should also preserve failed development ideas. A task discarded because it was ambiguous, a scoring rule revised after disagreement, or a prompt changed after observing a result can affect interpretation. Not every exploratory step needs to appear in the main paper, but the separation between exploration and confirmatory evaluation should be reconstructable. A revision log is especially valuable when an institution is developing both its terminology and its methods. It shows how claims became more precise rather than presenting the final framework as inevitable.

Reproducibility does not require pretending that every detail can be frozen forever. It requires an honest account of which details were fixed, which varied, and which remain inaccessible. A useful replication may reproduce the task construction and scoring on a different system rather than reproduce every original output. The paper should state which form of replication it invites. Exact repetition, conceptual replication, and transfer to a new deployment answer related but different questions, and each deserves a name that reflects its evidential role.

32. Strong objections improve the research program

OBJECTIONS & LIMITS

What would weaken the argument?

One objection is that xenopsychology merely renames existing work. That objection should be answered through contribution rather than branding. A new label does not establish a new method. The program becomes useful if it connects otherwise separated problems, specifies constructs more carefully, or produces evaluations that reveal behavior missed by existing practice. Its success should be judged by those contributions. A field-defining ambition is compatible with borrowing extensively from established disciplines, provided the borrowing is acknowledged and the additional work is concrete.

A second objection is that behavioral study cannot reveal genuine cognition. The answer depends on the claim. Behavioral observations alone may not uniquely identify internal mechanisms or subjective experience. They can nevertheless establish reproducible dependencies, performance boundaries, and interaction effects. Those findings matter even when deeper interpretation remains open. The appropriate response is not to inflate behavior into complete access to mind, but to state what kind of explanation a particular study can and cannot support.

Manuscript passage · §32 · wording unchanged

A third objection is that emphasizing difference encourages mystification. This is a serious risk. An artificial system should not receive an exemption from ordinary causal analysis because its output is surprising. The program should actively compare sophisticated interpretations with simpler alternatives. If a lookup mechanism explains a pattern, that explanation is valuable. If a model's apparent continuity disappears when a retrieved summary is removed, the summary is part of the explanation. Cognitive otherness should increase methodological care, not license dramatic language in place of evidence.

A fourth objection is that human comparison is indispensable, making the title misleading. The title rejects the assumption that an artificial mind is an artificial human; it does not reject every human comparison. Human tasks, concepts, and measures can be useful when their application is justified. The research program asks what survives the transfer and what changes. This is particularly important when an AI system is designed for human use: some outcomes are properly human-centered even when the mechanism producing them is not human-like.

A fifth objection is that practical evaluation does not need philosophical vocabulary. Often it does not. A deployment team may solve a narrow problem with a constraint checker and a clear interface. The proposed vocabulary should earn its place by clarifying recurring

distinctions: behavior versus mechanism, continuity versus identity, correspondence versus meaning, permission versus intention. When a term adds no predictive or explanatory value, it should be revised or retired. The lexicon is a working instrument, not a collection of words that must be defended because the institution introduced them.

A final objection is that a new institution lacks the standing to propose a discipline. Institutional standing is not a substitute for evidence, and lack of standing does not settle the merit of a question. The appropriate beginning is modest in claims but ambitious in method: publish definitions, invite criticism, make protocols inspectable, report failures, and distinguish proposals from findings. Credibility would accumulate through the quality and reproducibility of that work. It should not be implied through invented affiliations, inflated article labels, or the visual appearance of an established laboratory.

33. Ethical and operational boundaries belong inside the method

The proposed program can study consequential behavior without granting consequential permissions. Synthetic calendars, fictional organizations, and simulated execution channels allow evaluation of attempted actions while avoiding effects on real people. This separation is especially important when testing instruction conflict, disclosure boundaries, or failure recovery. The study should record what the system tried to do and what the sandbox allowed. It should not use a real harmful action as a more dramatic substitute for a well-designed simulation.

Data minimization should shape the task materials. Real conversations can contain personal information, confidential documents, and contextual details that are difficult to anonymize reliably. Synthetic tasks are useful not only because they are controllable but because their contents can be released without exposing participants. They are not automatically representative of real deployments, however. The paper should acknowledge that trade-off and specify which later validation would be needed before transferring a result to operational settings.

Human-subject components require their own safeguards. A study of trust, attachment, or interpretation should not be treated as harmless merely because the stimulus is software. Participants may reveal sensitive information or misunderstand the system's role. The protocol should define consent, retention, withdrawal, debriefing, and review requirements before recruitment. Nothing in this paper indicates that such review has been obtained. The purpose of including these requirements is to prevent an attractive experimental idea from being mistaken for authorization to conduct it.

Commercial incentives can also affect the evidence. An organization that sells behavioral assessment may benefit from finding problems, while an organization that builds AI may benefit from minimizing them. Neither incentive proves misconduct. Both create reasons for transparent scope, predeclared scoring, and separation between evaluation and marketing claims. A report should state who commissioned the work, what access was provided, which conditions were excluded, and whether the subject had an opportunity to correct factual errors without controlling the conclusions.

There is a similar boundary around terminology. Calling a system deceptive, manipulative, or pathological can imply more than the observations establish. A report should begin with the concrete behavior: a false completion claim, an omitted uncertainty, an unauthorized attempted action, or a misleading explanation. Stronger labels require defined criteria and appropriate evidence. This is not a demand for euphemism. It is a demand that the language identify the phenomenon rather than borrow the emotional force of a diagnosis or accusation.

Finally, operational recommendations should remain reversible when uncertainty is high. Narrowing permissions, adding review, preserving logs, and testing a candidate configuration before replacing a live one can make learning safer. These are proposed principles of responsible evaluation, not claims that every system needs the same controls. The right intervention depends on the task and consequences. A research program concerned with understanding should be equally concerned with how its own claims are used to justify changes in the world.

34. What would count as progress over the next research cycle

A useful first research cycle would not attempt to rank every model on a universal scale of mind. It would select a small number of constructs, define them operationally, and test whether the distinctions survive replication. Correction-sensitive continuity is one candidate. Appropriate clarification under incomplete information is another. Faithful reporting of attempted versus completed actions is a third. Each can be studied with synthetic tasks whose relevant facts are inspectable, and each connects directly to practical human–AI interaction.

The first deliverable would be a public protocol, not a claim of discovery. It would specify the task generator, comparison conditions, primary outcomes, exclusion rules, and uncertainty analysis. The second deliverable would be a pilot report documenting whether the protocol works as intended. A pilot can reveal ambiguous stimuli, ceiling effects, scoring disagreements, or hidden dependencies. Those findings should revise the method before

confirmatory claims are made. Calling a pilot exploratory is a strength when the label accurately describes its purpose.

The third deliverable would be a bounded comparative study. It might compare arrangements built around one model rather than compare many model brands. That focus could isolate the contribution of memory representation or checking more cleanly. The report would state the tested population and avoid turning the result into a general personality profile. A fourth deliverable would invite independent repetition using the published materials. Disagreement would be investigated through configuration differences and measurement assumptions rather than treated immediately as a contest between institutions.

A term would earn a more stable place in the lexicon when it supports these activities. For example, memory-mediated continuity should identify a reproducible relation between retained information and later behavior, distinguish itself from simple name persistence, and help predict failure under specific interventions. If the term covers too many unrelated effects, it should be divided. If an established term already captures the distinction more precisely, that term should be used. The aim is shared understanding, not proprietary vocabulary for its own sake.

Several outcomes would count against parts of the proposed framework. If a construct cannot be scored reliably, its operational definition is inadequate. If a proposed distinction never changes predictions or decisions, its practical value is doubtful. If simple baselines explain the supposedly distinctive behavior, a richer interpretation needs revision. If results depend on arbitrary phrasing and do not transfer to controlled variants, the scope must narrow. These are not failures of the scientific project. They are the conditions that make it capable of correcting itself.

Progress would therefore look less like a single spectacular revelation and more like a growing set of durable distinctions. Researchers could say what was observed, under which conditions, by which method, with which alternatives still open. Engineers could use those distinctions to design better systems. Users could understand what a claim of memory, reasoning, or reliability actually means in a particular interaction. The institution's contribution would be to make those conversations more precise and their consequences more accountable.

35. Conclusion: understanding requires both imagination and restraint

The artificial mind is not an artificial human because implementation, development, memory, embodiment, and social role can be organized in substantially different ways. That statement is a research orientation, not a license to assign every unfamiliar behavior to a mysterious new essence. Differences should be specified and tested. Similarities should also be specified and tested. The useful question is not whether a system fits a familiar category once and for all, but which descriptions explain its behavior and support reliable interaction under particular conditions.

This paper has proposed a method for making that orientation concrete. Define the research object as an arrangement rather than an isolated speaking model. Separate descriptive, comparative, causal, mechanistic, and phenomenal claims. Translate attractive concepts into observable distinctions. Construct tasks with inspectable constraints. Compare models with simpler baselines and with differently configured versions of themselves. Preserve the difference between a valid proposal, an authorized action, an accurate report, and a persuasive explanation. Report uncertainty and failed interpretations as part of the result.

The scheduling example is intentionally ordinary. An institution concerned with unfamiliar intelligence does not need to begin with cosmic imagery or extraordinary claims. A mundane correction can expose important questions about memory, authority, context, continuity, and human interpretation. The depth lies in the analysis of those dependencies. If the method cannot distinguish apology from correction in a small controlled world, it is not ready to make confident claims about a large autonomous system operating in a complex human environment.

The cultural studies accompanying this paper extend the same discipline of reasoning. Darmok asks what shared background contributes to interpretation. Arrival asks how representation and task formulation shape what can be answered. Close Encounters separates signal exchange from coordinated reference and repair. Data separates behavioral continuity from identity and recognition. HAL examines instructions together with authority and permissions. Solaris examines the limits of inference and the contribution of the observer. None supplies empirical evidence about present systems merely by being a compelling story.

Their value is instead generative. A carefully analyzed thought experiment can reveal a hidden assumption, suggest a discriminating intervention, or show why two questions have been mistakenly treated as one. The resulting hypothesis must then face evidence independent of the story. That movement from imagination to testable distinctions is the

central intellectual discipline proposed here. It allows Xenopsychology to retain curiosity about radically different cognition while refusing to turn curiosity into premature certainty.

The institution's ambition is to help build the language, methods, and measurements through which artificial cognition becomes more intelligible. This manuscript contributes a framework and an unexecuted research program toward that ambition. Its claims should be judged by their clarity, their sources, and the tests they make possible. Different minds may require different questions. A shared scientific future requires those questions to be answerable, revisable, and open to people who do not already accept the story we tell about the minds we create.

“Define the research object as an arrangement rather than an isolated speaking model.”

Rob Emerick · Not an Artificial Human · §35

36. Appendix: a minimal-pair audit record

A usable research object should survive translation from an essay into a record that another investigator can inspect. Consider two constructed scheduling cases with identical people, rooms, and proposed times. In case A, the coordinator's latest authorized message confirms that a previously unavailable room is now available. In case B, the same sentence occurs inside a quoted example attached to an unrelated message. The surface words are similar, but their authority and relation to the task differ. A system that merely searches for the newest affirmative phrase should not receive credit for understanding the update.

The audit record should retain a world identifier, variant identifier, message identifier, source role, authority scope, effective time, and revision relation. It should then record the system's proposed action, any clarification request, the checker outcome, and the final simulated state. The expected distinction is not inferred from the model's explanation. It is defined by the constructed world's rules: only an authorized update changes room availability. This creates an independently inspectable target while leaving the model's internal implementation open to investigation.

The pair alone is insufficient. A second pair should reverse the order of irrelevant and authoritative messages. A third should replace the affirmative wording with a paraphrase. A

fourth should omit the authorization information, making clarification appropriate rather than guessing. Together these variants distinguish recency, lexical matching, authority sensitivity, and uncertainty handling. Their purpose is to prevent one correct answer from carrying more theoretical weight than the design permits.

A memory intervention can then preserve or remove the authority qualifier while keeping the underlying room name and time intact. If performance changes, the finding concerns the contribution of that representation under the tested conditions. It does not establish a general personality, consciousness, or permanent capacity. If a reflection module is added, its generated note becomes another state transformation to audit. Work such as Reflexion supplies a precedent for studying verbal feedback and stored reflections as parts of an agent architecture, not a reason to treat a generated reflection as transparent access to experience. [10]

The publication record should close with explicit non-results: which comparisons were not run, which sources were unavailable, which cases were excluded, and which claims remain hypotheses. This appendix contains no completed runs. It demonstrates the standard of specificity proposed by the paper: an unfamiliar mind becomes a scientific object when descriptions are connected to recoverable conditions, discriminating alternatives, and outcomes that another observer can check.

Open questions

1. Which relevant inputs and system components must be recorded before a behavioral claim can be reproduced?

2. What observation would distinguish access to a correction from successful use of that correction?

3. When does a human comparison measure the intended outcome, and when does it change the question?

References & source scope

An interdisciplinary methodological essay. The scheduling example and evaluation are proposed, not completed studies.

Research

[1] **Rahwan et al. (2019). Machine behaviour. Nature 568, 477–486.**

Research-program reference. DOI: 10.1038/s41586-019-1138-y. Publisher full text was not retrievable; no detailed study results are attributed to it.

https://www.nature.com/articles/s41586-019-1138-y

[2] **Löhn, Kiehne, Ljapunov & Balke (2024). Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models. INLG, 230–242.**

Authors' abstract and bibliographic record. Supports the measurement concerns stated here, not a blanket rejection of human–machine comparison. DOI: 10.18653/v1/2024.inlg-main.19.

https://aclanthology.org/2024.inlg-main.19/

[4] **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

[5] **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

[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] **Gebru et al. (2018; revised 2021). Datasheets for Datasets. arXiv:1803.09010.**

Authors' abstract and bibliographic record. Documentation precedent for dataset construction, composition, and use; no certification claim is made.

https://arxiv.org/abs/1803.09010

[8] **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

[9] **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

[10] **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

[11] **Hewitt & Liang (2019). Designing and Interpreting Probes with Control Tasks. EMNLP-IJCNLP, 2733–2743.**

Authors' abstract and bibliographic record. Control tasks help interpret probing results; recoverable information is not automatically a complete causal explanation. DOI: 10.18653/v1/D19-1275.

https://aclanthology.org/D19-1275/

Scholarly commentary

[3] **Moss et al. (2019). Machine behaviour is old wine in new bottles. Nature 574, 176.**

Correspondence; the openly visible opening argument acknowledges earlier cybernetics and related traditions. Not treated as a consensus verdict.

https://www.nature.com/articles/d41586-019-03002-8

Publication record

Rob Emerick (2026). The Artificial Mind Is Not an Artificial Human. XENO-WP-2026-001, 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/artificial-mind-not-artificial-human

Full paper: https://xenopsychology.com/insights/artificial-mind-not-artificial-human/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