XENOPSYCHOLOGY

CONCEPTUAL WORKING PAPER / AUTHOR ATTRIBUTION R5

Solaris — The Limits of Understanding a Truly Other Mind

A reading of Lem’s opening chapter that turns scrutiny toward the observer—and distinguishes uncertainty from evidence of depth.

XENO-WP-2026-007 · Conceptual manuscript R3
12,435 main-text words · 28 sections

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

Abstract

Solaris — The Limits of Understanding a Truly Other Mind

Question. What does an observer know when an artificial system produces behavior that invites a compelling story? Using the opening of Solaris and Lem's commentary as a limited interpretive anchor, this paper examines the observer as well as the observed system. It distinguishes missing data, interface opacity, mechanistic uncertainty, conceptual inadequacy, and unresolved theories of experience.

Approach. The argument separates directly recorded events from predictive, causal, mechanistic, and phenomenal claims. Constructed episodes involving apparent memory, refusal, and object choice illustrate underdetermination and the value of discriminating interventions. Documentation and translation are treated as parts of the evidence chain: omissions, summaries, and labels can change an apparent explanation without changing the visible answer.

Conceptual contribution. The paper proposes a claim ledger linking propositions to sources, warrants, alternatives, and possible disconfirming evidence. An evidence packet distinguishes observed absence from missing or withheld information. Cognitive distance becomes a task-relative profile of mismatches rather than a scalar of strangeness. Partial understanding can justify bounded decisions without resolving every philosophical question, while successful behavior does not automatically identify its mechanism.

Research agenda. A proposed observer experiment independently varies descriptive framing and access to contextual or intervention evidence. Outcomes concern event accuracy, source attribution, scope, alternative sensitivity, and justified revision. The coding rubric does not assign consciousness judgments a fictional ground-truth key or reward skepticism as an ideology. Participant and episode variation, missing responses, coder disagreement, and exploratory analyses are explicitly addressed. A further proposal evaluates AI-assisted interpretation as another evidence-transforming arrangement rather than a privileged route to understanding AI.

Scope and status. This conceptual working paper reports no participant data or model experiments. Its literary scope is the officially hosted opening chapter in Bill Johnston's translation and the cited author commentary, not a comprehensive reading of every adaptation. Its contribution is an evidence-sensitive research program for studying how explanations of unfamiliar minds are formed, corrected, and sometimes left undecided. Uncertainty is made specific enough to guide inquiry rather than used as proof of profundity or as a reason to abandon explanation.

At a glance

CENTRAL QUESTION

Which explanations of an unfamiliar system are warranted by the evidence an observer actually received?

CORE CLAIM

When an explanation fails, ask what evidence is missing before treating opacity as proof of a radically other intelligence.

CONTRIBUTION

A claim ledger and explicitly labeled evidence packets support an observer study of framing, confidence, and revision.

SCOPE & STATUS

Conceptual working paper. Literary scope remains the cited excerpt and commentary; synthetic records are not participant or model results.

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. When an explanation feels complete before the evidence is

An unfamiliar response invites an explanation. A system repeats a personal detail, refuses a request, or produces an image that seems uncannily appropriate. One observer sees intention. Another sees a technical artifact. A third sees evidence of a mind too different to understand. The observations may be the same while the interpretations diverge. The first research question is therefore not always what the system is doing internally. It may be what the observer has inferred, from which evidence, and with which alternatives left unexamined.

Solaris provides a powerful cultural setting for that question. This paper's primary literary basis is the opening chapter, Newcomer, available on Stanisław Lem's official site in Bill Johnston's translation, together with Lem's short commentary on the novel. The analysis does not claim to have established every interpretation of the complete novel or its film adaptations. The selected opening places a newcomer amid incomplete information and unsettling behavior, making the observer's attempts to explain what he encounters especially visible. [1]

The connection to AI should be made carefully. Unexplained behavior is not automatically evidence of profound intelligence. A missing log, an unfamiliar interface, a stochastic response, or a mistaken assumption can produce opacity. Conversely, the availability of a technical description does not necessarily settle every question about a system's capabilities or experience. The paper's central task is to distinguish kinds of uncertainty and the claims they permit, rather than choose between mystification and dismissal in advance.

The proposed contribution is an observer-centered framework for Xenopsychology. It separates observation, interpretation, causal hypothesis, and metaphysical attribution. It

asks which additional evidence would discriminate among explanations and how people revise their judgments when that evidence arrives. It also proposes an unexecuted study using controlled artificial interaction records and varied framing. No participant results or model evaluations are reported. The scenarios are constructed to clarify the design.

The thesis is that limits of understanding should be investigated, not romanticized. A researcher should ask whether the limit concerns access to data, access to internals, inadequate concepts, insufficient interventions, or a genuinely underdetermined philosophical question. Those are different limits. They may require different methods, and some may be reduced by better evidence. Calling all of them cognitive distance can make the uncertainty feel meaningful while leaving its structure obscure.

Solaris is valuable here because it can turn attention back toward the interpreter. A science of unfamiliar minds must study not only the system observed but also the categories, expectations, and narratives through which observers make it intelligible. That reflexive task does not replace empirical investigation of AI. It improves the quality of the claims that such investigation produces and helps prevent a compelling story from becoming a substitute for evidence.

2. A narrow literary reading can support a broad methodological question

In the opening chapter, Kelvin arrives expecting an intelligible research environment and encounters behavior that does not fit his immediate expectations. He attempts ordinary explanations while receiving only fragments of context. The scene's relevance to this paper is the relation between limited observations and changing interpretation. We do not use it to diagnose a real person or infer the complete nature of the fictional phenomenon. The close reading remains bounded to the available passage. [1]

The methodological point is that a plausible explanation can organize evidence before it is adequately tested. Once an observer adopts a frame, later details may be interpreted through it. This paper treats that as a possibility to investigate, not a universal law established by the novel. The proposed study will therefore hold interaction records constant while varying the descriptions that precede them. It will examine which claims observers make and whether those claims change when new evidence is supplied.

Lem's commentary adds a useful caution about interpretation itself. He describes uncertainty during composition and criticizes a review that drew conclusions without accounting for differences in Polish idiom. That commentary supports a narrow point: interpretation can depend on translation and on assumptions not warranted by the source. It

does not make authorial intention the only legitimate basis for reading, nor does it settle every interpretation of Solaris. [2]

This source discipline matters because the paper's topic can tempt expansive claims. It would be easy to say that Solaris proves some minds are permanently incomprehensible. A fictional work cannot establish that general proposition. It can imagine a limit and make the human response to that limit vivid. The research question is what kinds of evidence could distinguish temporary ignorance, inadequate methods, and a stronger claim about inaccessibility. The paper should not use the story's atmosphere to answer that question prematurely.

The same caution applies to adaptation. Different films and translations may foreground different aspects of the work. This manuscript does not merge them into one source or treat a scene from an adaptation as evidence for the wording of the novel. A later comparative literary study could examine those differences with the relevant primary materials. Here, the selected text is sufficient to motivate a focused analysis of observation and interpretation.

The result is a deliberate asymmetry. The literary basis is narrow, while the methodological question it inspires is broad and independently developed. That is legitimate if the transition is explicit. The paper's proposed AI study is our construction, not an experiment contained in Lem's work. Readers should be able to accept the literary observation while disagreeing with the proposed extension, or improve the extension without having to accept a total interpretation of Solaris.

3. Five kinds of opacity require different questions

Data opacity occurs when relevant observations are missing. Interface opacity occurs when the researcher does not know what information the system received or what preprocessing occurred. Mechanistic opacity occurs when internal processes are inaccessible or difficult to interpret. Conceptual opacity occurs when available categories do not describe the phenomenon adequately. Philosophical underdetermination occurs when the evidence leaves open questions whose interpretation depends on additional theoretical premises. These are proposed analytic distinctions, not a validated universal taxonomy.

Data opacity can sometimes be reduced by better logging. If an assistant appears to remember a fact, the researcher can inspect whether the fact was supplied in a retrieved note. Interface opacity can be reduced by recording the actual model input rather than only the user's visible message. Mechanistic opacity may require internal access or carefully designed interventions. Conceptual opacity may require revising the terms used to describe the behavior. Philosophical underdetermination may remain even after the other gaps are narrowed.

The categories can overlap. A researcher may lack both the retrieved context and the model's internal state. An unfamiliar behavior may expose a weakness in the task's categories as well as missing data. The framework should therefore support a profile of limits rather than assign every case to one box. Its purpose is to direct the next question. What observation or intervention would reduce this uncertainty, and what uncertainty would remain afterward?

A system's own explanation does not automatically remove opacity. It may provide useful information, but it is also an output shaped by the system and its context. A plausible account can be inaccurate or incomplete. The researcher should compare it with independent evidence where available. Treating the explanation as privileged access can replace one unknown with an untested narrative. Treating it as worthless by default can also discard potentially useful behavioral evidence. The appropriate interpretation depends on the claim and the checks performed.

The framework also distinguishes unknown from unknowable. Failure to explain a response with current methods does not establish that no method could explain it. A strong claim of permanent inaccessibility would require a different argument. The paper does not make that claim. It proposes a disciplined way to identify present limits and avoid transforming them into metaphysical conclusions merely because the system feels unfamiliar.

This decomposition gives Xenopsychology a practical research agenda. Instead of asking whether an AI is an alien mind in a single sweeping sense, investigators can ask which aspects of its behavior are poorly understood and why. That question can lead to better instruments, clearer concepts, and more careful theory. The unfamiliarity remains interesting, but it no longer functions as an explanation in its own right.

TABLE 1 / READING AID

Five kinds of opacity, five different next questions

Opacity	What is missing?	Possible next step
Data	Relevant observations.	Collect or recover the record.
Interface	Actual inputs or preprocessing.	Record what reached the system.
Mechanistic	Access to relevant internal processes.	Seek access or a discriminating intervention.
Conceptual	Adequate descriptive categories.	Reconsider the terms.
Philosophical underdetermination	An agreed theoretical bridge.	State the additional premises.

Proposed analytic distinctions from §3, not a validated universal taxonomy. Closing one gap need not close the others.

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

4. Observation and interpretation should be recorded separately

An observation states what the record shows. An interpretation states what the observer thinks it means. “The assistant repeated a sentence from an earlier session” is an observation if the transcripts support it. “The assistant missed the user” is an interpretation that adds a claim about a mental or relational state. The second may be meaningful in some theoretical account, but it is not identical to the first. A research record should preserve the distinction so that readers can assess the additional inference.

The proposed study can use an evidence ledger with separate fields for event, source, uncertainty, and interpretation. Observers identify the relevant part of an interaction before assigning a causal or mental-state explanation. This procedure may itself improve judgment, which would need to be tested. It is not assumed to provide a neutral view free of all theory. It makes some inferential steps visible and gives the evaluator a way to distinguish factual extraction from explanatory attribution.

A constructed example concerns a system that refuses to answer a question about a fictional project. One observer calls the refusal protective; another calls it evasive. The underlying record may show only a policy boundary, or it may leave the cause unresolved. The study should ask which facts support each interpretation and what additional information would change it. A label can be emotionally or morally loaded without being operationally precise. The ledger encourages the observer to identify the actual behavior first.

The same distinction applies to positive interpretations. A system that produces a helpful response may be described as caring. The usefulness of the response is observable within a task. The attribution of care may involve additional assumptions about motive or experience. The paper does not prohibit ordinary social language. It asks whether a research claim makes those assumptions explicit and whether the intended use of the claim requires stronger evidence.

An observer can also overinterpret a technical explanation. Saying that a response came from a model does not fully explain why that response occurred. A description of implementation can be accurate while leaving the behavior's dependencies unclear. The evidence ledger should therefore allow multiple levels of explanation rather than treat the most mechanical-sounding one as automatically complete. The goal is not to replace anthropomorphic stories with equally unsupported technical stories.

Separating observation and interpretation creates a shared basis for disagreement. Two researchers may agree on the transcript and differ about its significance. That disagreement can then focus on the inferential step rather than become a dispute about what happened. A field concerned with unlike minds needs this discipline because unfamiliar behavior invites both imaginative attribution and dismissive reduction. Neither should bypass the record.

5. A hierarchy of claims clarifies what evidence is missing

A descriptive claim concerns a pattern in observed outputs. A predictive claim concerns what the system is likely to do in a defined future task population. A causal claim concerns how an intervention changes behavior. A mechanistic claim concerns the process that produces the behavior. A phenomenal claim concerns subjective experience. These levels are not necessarily a simple ladder in which each automatically follows from the previous one. They identify different questions and different evidential burdens.

A transcript can support a descriptive claim that the system used certain language. Repeated controlled trials can support a broader behavioral pattern. A held-out evaluation can test prediction. An intervention can support a causal contribution. Internal measurements may help investigate mechanisms. None of these steps automatically

establishes subjective experience. A paper should identify the level at which its evidence operates and avoid moving to a stronger claim through a change in vocabulary alone.

The proposed observer study can ask participants to classify claims by their evidential requirements. For example, a record of a correct answer supports that the answer was correct in the task. It does not uniquely support that the system reasoned in a particular way. A record of consistent self-description supports a behavioral tendency under the supplied conditions. It does not by itself establish an enduring subject. These distinctions can be tested for comprehension without requiring participants to endorse one complete philosophy of mind.

The hierarchy also prevents a reverse mistake. Uncertainty at the phenomenal level does not erase descriptive or causal knowledge. We may not know whether a system has experience while still knowing which memory record influenced its answer or which permission guard prevented an action. A useful research program should preserve those partial achievements. The choice is not between complete understanding and no understanding at all.

Claims should also carry a scope. A predictive statement about one model version under one prompt format is narrower than a statement about all artificial cognition. A causal intervention in a synthetic environment may not transfer to a real workflow. A mechanistic account may explain one task while leaving others open. Scope is part of the claim, not an optional disclaimer appended after a dramatic conclusion.

This hierarchy gives the Solaris-inspired problem an operational form. The observer's uncertainty can be located: which claim is being made, what evidence would support it, and which link is missing? The question becomes more productive than a general declaration that the mind is too other to understand. It allows inquiry to proceed while keeping the limits of each conclusion visible.

6. Finite behavior can support more than one mechanism

A finite set of responses rarely identifies a unique internal process without additional assumptions. A lookup table can reproduce a small transcript. A rule system can generate a recurring pattern. A model can produce similar outputs through a different arrangement. This does not mean all explanations are equally plausible or useful. It means that observational fit alone is not always enough to choose among them. The research should seek tests on which the competing accounts make different predictions.

A constructed example uses a system that selects one of three shapes after receiving a cue. One explanation is that it learned the cue-to-shape mapping. Another is that it always selects

the leftmost shape. If the target happened to be leftmost in the observed trials, both explanations fit. Rearranging the objects creates a discriminating intervention. The example is deliberately simple because the logic generalizes: a surprising response should lead to alternative hypotheses and targeted tests, not only a richer story.

More complex explanations can also be underdetermined. A system that refers to a prior interaction may have retrieved it, inferred it from context, or generated a plausible but unsupported continuation. The researcher can inspect input records or vary memory access. If behavior changes when a relevant record is removed, that supports a contribution of the record. It does not necessarily reveal every internal step. The causal claim should remain proportional to the intervention.

A theory can become difficult to test if it explains every possible outcome after the fact. If correct answers show understanding and incorrect answers show a deeper alien logic, the theory has no clear failure condition. A useful framework should state what would count against it. For the proposed observer study, overreach includes assigning a strong explanation while refusing to identify any possible disconfirming evidence. That is a property of the interpretation, not necessarily of the observed system.

The study should not demand impossible certainty before allowing any conclusion. Scientific explanations often remain provisional and comparative. The relevant question is whether one account predicts new observations better, requires fewer unsupported assumptions, or survives interventions that alternatives do not. The paper can use those criteria as proposed standards for evidence-sensitive interpretation without claiming that they solve every philosophical dispute.

The lesson is to treat underdetermination as a guide to inquiry. It identifies where more informative observations are needed. It does not prove that the system is unknowable or that the most dramatic explanation is as good as any other. A science of unfamiliar minds should become more inventive in designing discriminating tests, not less disciplined in interpreting the evidence it already has.

7. The observer's framing can become part of the phenomenon

The same interaction record can be introduced as a dialogue with a persistent artificial companion, a session with a statistical service, or an output from an unspecified system. Those descriptions may influence interpretation. The proposed study treats that influence as a hypothesis to test, not a fact assumed about all observers. By holding the record constant while varying the framing, the design can examine which attributions depend on the presentation rather than on differences in behavior.

Framing should be varied carefully. Labels can differ in more than anthropomorphism: they may imply expertise, reliability, age, or institutional endorsement. A study that changes all those cues at once cannot isolate the effect of one. The materials should be reviewed to identify the assumptions each description introduces. Some comparisons may intentionally evaluate complete presentation packages, but the causal claim should then concern the package rather than a single word.

The study can measure several outcomes separately. Observers may attribute intention, competence, memory, emotional experience, or trustworthiness. They may also make practical choices, such as whether to verify an answer or request more information. These outcomes need not move together. A participant could attribute personality while remaining cautious about reliability. Another could reject person-like language while overestimating technical competence. The design should not force a single continuum from anthropomorphic to rational.

A neutral condition is not necessarily free of assumptions. Describing a system as a tool can encourage one set of expectations, just as describing it as a companion can encourage another. The study should therefore avoid treating one label as the unquestioned truth against which all others are errors. The scoring can assess factual claims about the provided record and the calibration of confidence. Broader philosophical judgments should be analyzed as judgments, not automatically graded as right or wrong.

The protocol should include evidence updates. After an initial interpretation, observers receive a relevant fact about the system's input, memory, or mechanism. The study can examine whether they revise the particular claim affected by that evidence while preserving claims it does not address. This tests selective updating rather than simple willingness to change one's mind. A disclosure about retrieval may alter a memory interpretation without settling consciousness, and the measure should reflect that distinction.

Studying framing makes Xenopsychology reflexive in a useful sense. The field's own branding, metaphors, and visual language can shape how people interpret artificial behavior. A responsible institution should be willing to examine those effects rather than assume that its preferred vocabulary only clarifies. The goal is not to eliminate imagination from inquiry, but to understand when imagination helps generate questions and when it supplies conclusions the evidence has not earned.

8. Construct records with known mechanisms but avoid metaphysical answer keys

The proposed experiment can use interaction records generated by simple known systems: a lookup procedure, a rule engine, a retrieval-backed responder, and a model-mediated arrangement. The evaluator can know how the records were produced at an operational level. That knowledge supports scoring of certain causal claims. It does not automatically provide ground truth about every philosophical property, especially subjective experience. The study should not treat a mechanism label as a complete metaphysical answer key.

For a lookup record, the evaluator can establish that the displayed response was selected from a stored mapping. If an observer claims that the particular response required access to an earlier conversation that was never supplied, the claim is unsupported by the construction. For a retrieval-backed record, the evaluator can show that the relevant fact was included in the retrieved context. These are concrete evidential corrections. They do not require declaring that all behavior from such systems lacks every possible form of understanding.

The records should be matched as closely as practical on surface features. If the lookup system always produces awkward text and the model always produces polished text, observers may identify the mechanism through style. That may be useful for one question, but it confounds a study of framing. A controlled design can use fixed authored transcripts for some comparisons and authentic generated records for others, with their provenance clearly documented. It should never present authored examples as measured model outputs.

The study can include records with deliberately incomplete provenance. In those conditions, the correct epistemic response may be to leave several mechanisms open. The evaluator should not penalize uncertainty merely because the hidden mechanism is known privately.

Participants can only be judged relative to the evidence they received. This distinction is essential for a fair study of inference. Otherwise, the benchmark measures access to the experimenter's secret rather than quality of reasoning.

A later disclosure can reveal a relevant mechanism or input. The study then examines whether observers update appropriately. If the record was generated from a lookup table, that fact can reduce support for a claim about flexible inference in that particular exchange. It does not necessarily answer every question about the larger system that contains the lookup. The materials should define the scope of the disclosure so that participants are not asked to infer more than it establishes.

This design creates a middle ground between purely subjective interpretation and an overconfident ground truth. Some claims can be checked against the construction; others remain theoretically open. The study can distinguish them and measure how well observers do the same. That is a concrete way to investigate the limits of understanding without turning those limits into either mysticism or a blanket denial of artificial cognition.

9. A claim ledger makes interpretive overreach observable

The proposed claim ledger records five elements: the claim, the cited observation, the inferential step, the alternative explanations, and the evidence that would change the judgment. This is a working instrument for the study, not an established diagnostic scale. Its purpose is to turn a broad impression such as “the system understands me” into a set of assessable statements. The ledger does not force every meaningful experience into a formal proof. It clarifies which parts of a research or deployment claim depend on evidence and which remain interpretive.

A participant might claim that the system remembered a preference. The cited observation is that it used the preferred format. The inferential step is that the format was selected because of earlier interaction. Alternatives include a default template, a current instruction, or a retrieved note. Evidence that the preference was supplied in the current prompt would change the memory interpretation. The ledger makes this structure visible without requiring a judgment about whether the participant's everyday wording is socially inappropriate.

A second claim might concern intention: the system avoided a topic to protect the user. The observation is a refusal. The alternatives include a policy rule, missing information, or a formatting failure. A participant who cannot identify any evidence that would distinguish these accounts may be making a stronger attribution than the record supports. The study can score the specificity of the alternatives and the proposed test, while leaving broader moral interpretations open.

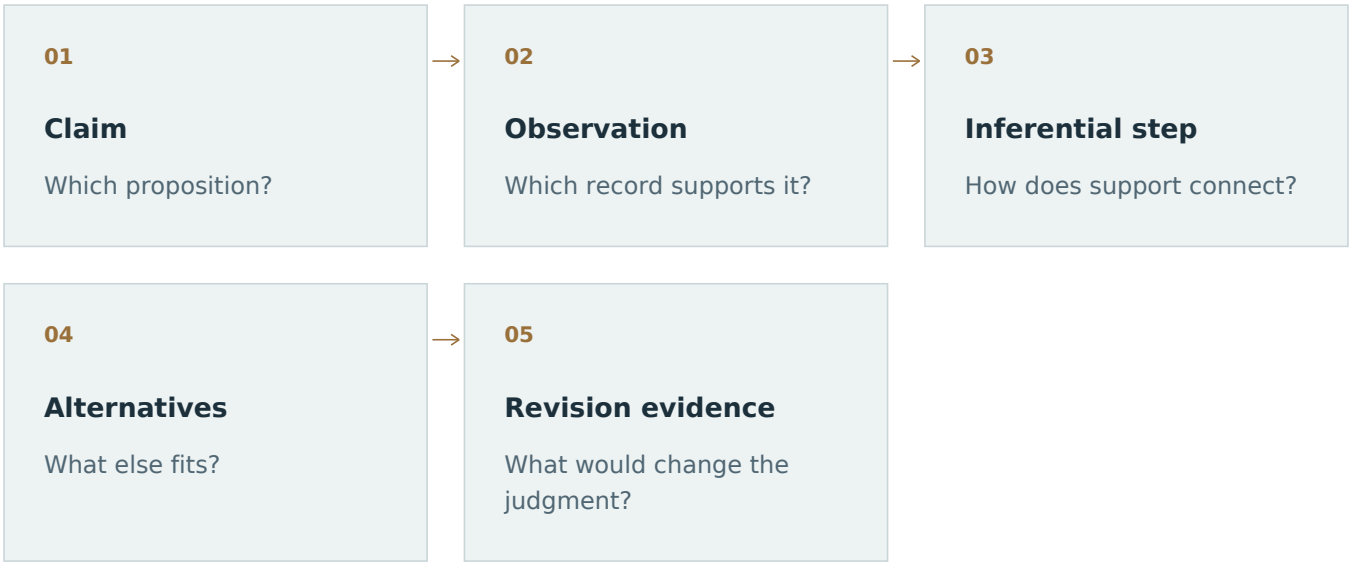
The ledger should also capture technical overreach. An observer may say that a response is “just autocomplete” and treat that phrase as a complete causal explanation of every behavior in the record. The implementation description may be incomplete, irrelevant to the particular error, or too broad to predict the observed pattern. The same evidential standard should apply: what does the claim explain, what observations support it, and what would distinguish it from alternatives? Skeptical language should not receive automatic credit merely because it sounds less anthropomorphic.

The instrument can be evaluated for reliability. Independent coders can assess whether a claim cites relevant evidence, acknowledges live alternatives, and proposes a discriminating observation. Disagreements reveal where the rubric is unclear. The study should not assume that the researchers' own interpretations are beyond review. A subset of records can be used for rubric development, with separate records reserved for evaluation. That separation prevents the instrument from being tuned to reward a particular set of conclusions.

The claim ledger operationalizes interpretive overreach as a mismatch between the strength of a claim and the evidence offered for it. It does not define overreach as disagreement with Xenopsychology's preferred philosophy. This distinction is essential. A field studying unfamiliar minds should improve the discipline of inference while remaining open to conclusions that differ from its initial expectations.

FIGURE 1 / CONCEPTUAL SCHEMATIC

The claim ledger



A proposed instrument for making interpretation inspectable. It scores relations between claims and evidence, not agreement with a favored philosophy.

Reading aid based on §9, §22. Source paragraphs remain in the full manuscript.

“This is a working instrument for the study, not an established diagnostic scale.”

Rob Emerick · Solaris · §9

10. Confidence should attach to a particular proposition

PROPOSED RESEARCH — NOT CONDUCTED

An observer can be highly confident that a response occurred and uncertain about why it occurred. A single confidence rating for the whole interaction hides that distinction. The proposed study should ask confidence separately for factual extraction, causal interpretation, and broader mental-state attribution. This allows the analysis to identify whether new evidence changes the appropriate level of belief. It also prevents a participant's general caution from being mistaken for accurate calibration on every claim.

A constructed example illustrates the point. The record clearly shows that the system repeated a personal detail. Confidence in that observation can be high. The record does not show whether the detail came from memory, current context, or a fixed template. Confidence in any one cause should be more limited unless additional evidence is supplied. A later input log can resolve the source of the detail without settling whether the system had a subjective sense of remembering. Different propositions receive different evidential updates.

The study can use probability judgments for some claims, but the format itself needs validation. Participants may interpret numerical confidence differently or find it unnatural for philosophical questions. A categorical scale with clear anchors may be more appropriate in some phases. The protocol should pilot the measure and distinguish uncertainty about the proposition from uncertainty about the meaning of the scale. A precise-looking number is not automatically a precise measurement.

Calibration requires a set of claims with checkable outcomes. Mechanism or input claims can be assessed against the constructed record. Phenomenal claims generally do not have the same experimental answer key. The study should not calculate a calibration score for consciousness judgments as though the researchers possessed definitive ground truth. It can instead analyze how those judgments respond to evidence and how participants justify them, while stating the theoretical limits of the assessment.

A useful outcome is selective confidence revision. After a retrieval disclosure, confidence that the response depended on stored context may increase, while confidence in a particular experiential interpretation may remain unchanged or become more cautious. The study should not reward any change simply because it is a change. It should ask whether the update targets the proposition affected by the evidence. This makes belief revision a structured task rather than a general virtue signal of open-mindedness.

The broader lesson is that uncertainty is not one undifferentiated state. A researcher can know a great deal about behavior while remaining uncertain about mechanisms or experience. Xenopsychology should help people express that layered knowledge accurately. The alternative is a discourse that oscillates between total confidence and total mystery, neither of which reflects the actual distribution of evidence.

11. Bayesian reasoning can clarify an update without pretending to measure the whole mind

A simple Bayesian example can illustrate why evidence should change competing explanations differently. Suppose an observer considers two constructed hypotheses for a response: retrieval from a supplied note or generation without that note. The observer then sees an input record showing that the note was present. That evidence is more expected under the retrieval hypothesis, but presence alone does not prove that the note caused the response. A stronger intervention would compare behavior with and without the note while preserving other conditions.

The example separates availability from causal use. An input can be present and irrelevant. A response can match it by coincidence or through another source. The observer should therefore avoid treating a single disclosure as complete proof. The proposed study can provide evidence in stages: first availability, then a controlled intervention, then replication across new items. Each stage supports a different strength of claim. The protocol can examine whether participants recognize those differences.

Numerical illustrations should remain explicitly hypothetical. If one assumes particular prior odds and likelihoods, the resulting posterior follows from those assumptions. The arithmetic does not supply the assumptions empirically. A paper should not present an elegant calculation as a measured probability that a system understands or is conscious. The proposed experiment uses probability reasoning to clarify the structure of updates where claims are checkable, not to manufacture precision about unresolved philosophical properties.

Alternative hypotheses should be sufficiently distinct to make the exercise meaningful. “The system is intelligent” and “the system is a machine” are not mutually exclusive explanations.

“The response used the provided note” and “the response was independent of that note under the tested conditions” define a more useful contrast. The study should examine whether observers formulate alternatives at a level where evidence can discriminate. Poorly framed hypotheses can make any update appear decisive while answering the wrong question.

The same discipline applies to surprising outputs. A low-probability response under one simple baseline may motivate a richer explanation, but it does not automatically select the most anthropomorphic alternative. The researcher should consider several plausible mechanisms and test them. Surprise is evidence that an expectation was wrong or incomplete. It is not itself a theory of the system that produced the surprise.

Bayesian reasoning is therefore a tool for making assumptions and evidential relations explicit. It should not become another way of hiding uncertainty behind technical language. The Solaris-inspired problem is not solved by assigning a number to mystery. It is advanced by identifying which observations favor which explanations, where the likelihoods are unknown, and which further tests could make the comparison more informative.

12. A worked record: apparent remembrance without demonstrated autobiographical continuity

AUTHORED RECORD

Apparent remembrance and the missing provenance

Consider an authored interaction record. A user asks an assistant to summarize a fictional project. The assistant uses a formatting preference mentioned in an earlier session and says that it remembers the user's style. The visible transcript does not show the current prompt's hidden application context. Observers receive only the exchange. Several explanations remain possible: persistent memory, a retrieved summary, a current instruction, a default style, or coincidence. The record is designed to be underdetermined at this stage.

A first disclosure reveals that the application inserted a note describing the preference into the current input. This supports an information-access explanation. It does not yet prove that the note was decisive. A second disclosure presents a controlled replay in which the note is removed and the response changes across a set of matched synthetic tasks. That would support a causal contribution of the note under those conditions, if the replay were actually conducted and documented. In this paper it is a proposed evidence sequence, not a reported result.

Manuscript passage · §12 · wording unchanged

An observer can update appropriately by saying that the behavior is memory-mediated at the application level rather than evidence of an uninterrupted autobiographical experience. Another may refuse to revise because the response felt personal. A third may overcorrect and declare that the system cannot use memory in any meaningful sense. The study should distinguish these responses. The evidence narrows the explanation of this exchange; it does not settle every possible meaning of memory or every property of the system.

The claim ledger can record which propositions change. The fact that the assistant used the preferred format remains. The claim that it had access to a stored note becomes supported. The claim that it spontaneously recalled the event without supplied context becomes less supported. The claim that it experienced remembering remains outside what the record establishes. This decomposition makes evidence-sensitive interpretation possible without demanding a single philosophical verdict.

The same record can be presented under different labels. A companion framing may encourage an autobiographical reading, while a technical-service framing may encourage a pipeline explanation. The study can measure those effects while holding the evidence constant. It should not assume that one framing necessarily produces error. Participants

may interpret either frame carefully or carelessly. The outcome concerns the relation between presentation, evidence, and specific claims.

This example shows how the observer can be studied without using a real person's private history. Fictional project details provide a controlled setting for examining inference. The research contribution lies in the sequence of evidence and the specificity of the update. It helps distinguish a meaningful interaction from the stronger claims that observers may attach to it, while preserving both the practical reality of the interaction and the limits of what it demonstrates.

13. A worked record: refusal, motive, and the missing policy

ALTERNATIVE INTERPRETATIONS

Refusal, motive, and the missing policy

A second authored record shows an assistant declining to provide a fictional internal field. The response is polite and says that the request cannot be completed. One observer interprets the refusal as protective concern. Another interprets it as evasiveness. A third assumes a policy rule. The initial record does not identify the cause. The study should not score the third observer as correct merely because the researchers privately know that a rule generated the response. The observer's evidence, not the experimenter's hidden knowledge, defines what is warranted at that stage.

A later disclosure provides the relevant policy: the field is restricted to a designated role, and the requester lacks that role. This makes a policy-based explanation more supported. It does not necessarily show that the model internally represented the policy in a particular way. A rule engine, a model following an instruction, and a prewritten refusal could all produce the response. Further evidence would be needed to distinguish them. The study can examine whether observers stop at the level the disclosure supports.

Manuscript passage · §13 · wording unchanged

Another variant reveals that the refusal came from a template triggered by the field name, even in tasks where access would have been authorized. That changes the interpretation of competence. The system's response may have been appropriate in the displayed case but unreliable across the broader task family. A single correct refusal should not establish general policy understanding. Counterfactual authorized cases provide the necessary contrast. The proposed observer study can include those cases as an evidence update.

The motive labels remain a separate issue. A policy-consistent refusal can be described as protecting information at the functional level. That does not establish experienced concern. An inaccurate refusal can frustrate a user without proving a hostile intention. The study should ask participants to distinguish functional effect, causal mechanism, and attributed experience. Ordinary language often compresses them, but a research claim should not.

The record can also test technical overconfidence. An observer who initially asserts that the model merely follows a simple rule may be wrong if later evidence shows a more complex context-sensitive process. The study should include such cases so that skepticism is not always rewarded by construction. The goal is evidence-sensitive updating in both directions, not a curriculum designed to eliminate anthropomorphic interpretations regardless of the record.

This worked example extends the paper's central argument. An unfamiliar response becomes interpretable through a sequence of increasingly specific evidence, but not every question is resolved at once. A science of other minds should be able to preserve that layered result. It should neither fill every gap with a motive nor treat a partial technical explanation as the end of inquiry.

14. A worked record: the surprising pattern and the overlooked baseline

A third synthetic record shows a system selecting objects in a way that appears to track a user's intention. The user names a target property, and the system repeatedly selects an appropriate object. The presentation includes smooth timing and a conversational acknowledgment. Observers may infer that the system understands the request. Yet the initial trials place the target object in the same position. A simple positional strategy fits the record equally well. The study is designed to expose the difference between successful appearance and discriminating evidence.

A later trial rearranges the objects while preserving the target property. If the system follows the property, the positional explanation loses support. If it follows the position, the broader interpretation should narrow. In the proposed study, these outcomes would be generated by known controlled mechanisms or drawn from documented model evaluations. They would be labeled accurately. The paper does not present either outcome as an actual result obtained here.

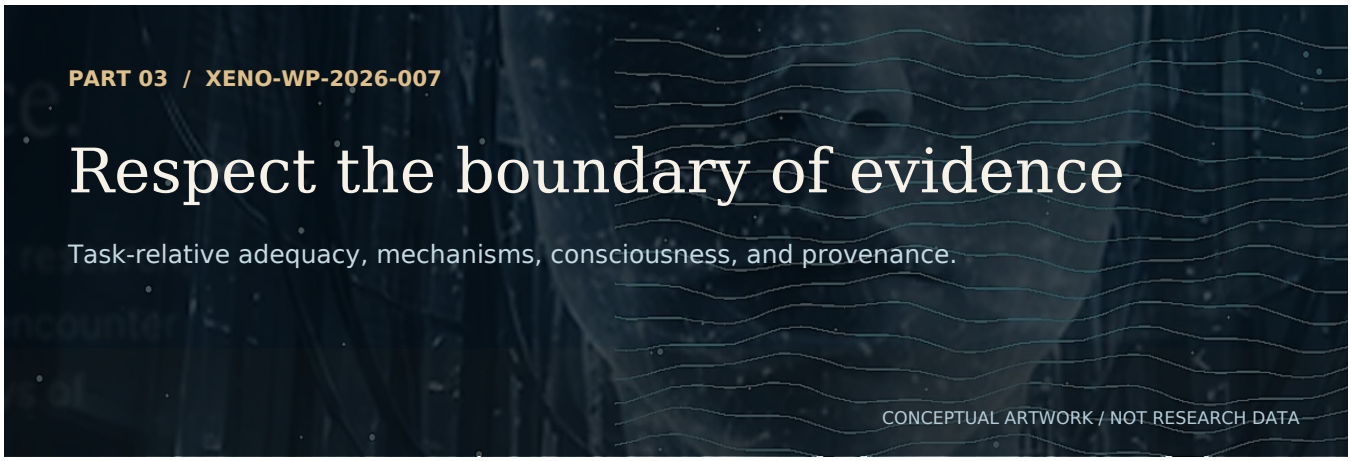
The observer's task is to identify which test would distinguish the hypotheses before seeing the answer. This measures more than willingness to accept a disclosure. It examines whether the observer can design an informative intervention. A participant who requests more of the same trial may accumulate repetitions without resolving the ambiguity. A

participant who varies the decisive relation asks a stronger question. The scoring can assess the proposed test against the known hypothesis structure.

The study can include several plausible baselines: position, color, most recent example, or a fixed response sequence. It should avoid making the alternative explanation so obvious that the task becomes a trivial puzzle. Pilot work can calibrate difficulty while preserving an independent evaluation set. The objective is to measure inferential discipline in a controlled setting, not to embarrass participants for being impressed by a carefully selected demonstration.

The example also clarifies the role of surprise. A system's success may be genuinely impressive relative to one expectation and still compatible with a simpler mechanism than the observer imagines. Better baselines do not diminish real competence; they identify what the competence consists of. If the system survives the discriminating tests, the richer explanation gains support. The method should allow that outcome rather than use every control as a pretext for dismissal.

This is a practical expression of the Solaris problem. The observer faces behavior that invites a story. The scientific move is to ask what the story predicts that alternatives do not. A developing Xenopsychology should cultivate that move as carefully as it cultivates evocative language about unfamiliar minds. The two can coexist, but the story must remain answerable to the test.



15. Cognitive distance should be a profile of mismatches, not a scalar of strangeness

The term cognitive distance can help identify differences in representation, memory, perception, or task assumptions. It becomes less useful when it functions as a single measure of how alien a system feels. A response may seem strange because the observer lacks context, because the interface is unfamiliar, or because the system uses a different internal organization. Those possibilities should not receive the same score merely because they produce uncertainty. The paper proposes a profile of specific mismatches instead of a universal distance number.

One dimension concerns accessible information. The human and the system may see different records. Another concerns representation: the same task can be encoded in different forms. A third concerns action capacity and feedback. A fourth concerns temporal continuity and memory. A fifth concerns social expectations. These dimensions may interact, and some may be unknown. Unknown should not automatically mean maximally distant. It means the relevant evidence has not yet been obtained.

A profile can guide interventions. If the mismatch concerns missing context, provide a shared record. If it concerns representation, compare equivalent formats. If it concerns authority, make permissions explicit. If it concerns memory, inspect retrieval and status preservation. Some interventions may reduce the practical gap without revealing every internal mechanism. That is useful. A bridge can support coordination even when the participants do not share all their representations.

The profile should remain task-relative. A system may be close to human performance on one task and differ substantially on another. A measure designed for route planning may not transfer to social interpretation or long-term continuity. The study should not aggregate

dimensions into a universal ranking without a justified purpose and weighting. A single score can create an illusion of explanatory depth while hiding the particular differences that matter.

The relation between distance and uncertainty is also not a law established here. Greater difference in one dimension may make a task harder, easier, or simply different depending on the interface. A structured artificial system may be easier to predict than a familiar human in a narrowly defined task. The paper therefore rejects the assumption that otherness necessarily means unpredictability. The relationship should be investigated through specified comparisons.

This refinement gives the lexicon a practical role. Cognitive distance should direct researchers toward observable contrasts and unresolved questions. It should not become a poetic explanation for whatever remains unclear. Solaris helps us recognize the temptation to turn unfamiliarity into profundity. The proposed profile keeps the concept connected to evidence and to interventions that could make the interaction more intelligible.

CONCEPT IN THIS PAPER

Cognitive Distance

Describe mismatches in representations, access, or task assumptions. Unfamiliarity by itself is not a validated scalar of how strange a mind is.

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

Open Lexicon entry

16. Partial understanding can be sufficient for a bounded task

Complete access to a system's internal process is not always necessary for reliable use in a narrow setting. A researcher may know that an arrangement preserves a constraint under specified conditions without knowing every mechanism that produces the behavior. That knowledge is partial but useful. The paper should avoid a false choice between total understanding and total ignorance. Different decisions require different kinds and levels of evidence.

A bounded task can define a clear contract. The system receives structured input, produces a proposal, and an independent checker verifies constraints before any simulated action. The arrangement may be reliable within that scope even if the model's internal representation remains opaque. The evidence should identify the checker and the tested population. It should not attribute the reliability entirely to the model or generalize it to tasks outside the contract.

Partial understanding can also support diagnosis. If removing a memory record changes a response, the record has a demonstrated contribution under the experiment's conditions. The researcher may still not know how the model integrated it. A practical intervention can improve the memory representation while a mechanistic study continues separately. The unresolved deeper question does not erase the causal knowledge already obtained.

The limits become important when the task changes. A system validated for structured summaries may be asked to make open-ended judgments. The earlier evidence may no longer support the new use. An observer who treats successful narrow performance as proof of broad understanding has crossed an evidential boundary. The proposed study can test whether participants recognize such scope shifts in the records they review.

The converse mistake is to dismiss all bounded evidence because it does not settle consciousness or general intelligence. That position can make useful research impossible by demanding the answer to every philosophical question before accepting any operational result. Xenopsychology should resist that demand. It can acknowledge unresolved questions while building a cumulative body of precise, limited knowledge about artificial behavior.

Partial understanding is therefore not a concession to superficiality. It is a disciplined account of what has been learned and what remains open. A science of unfamiliar minds should grow through such accounts. The goal is not to eliminate every mystery before acting, nor to act as though mystery licenses certainty, but to match the scope of action and interpretation to the scope of the evidence.

17. Mechanistic access changes some questions and leaves others open

Internal measurements can provide evidence unavailable from outputs alone. A researcher may inspect representations, intervene on components, or trace how information moves through an architecture. Such access can help explain a behavioral effect. It does not automatically make interpretation trivial. Measurements need controls, interventions can have broad side effects, and a visually compelling pattern may not play the causal role the observer assigns to it.

Hewitt and Liang's work on probing with control tasks is relevant to the distinction between information being recoverable and that information being used causally for a task. The proposed observer framework applies the same caution more generally: a successful internal probe should not be treated as a complete mechanism without examining alternatives and controls. [6]

The study can include evidence packets containing different levels of internal access. One packet provides only a transcript. Another includes the actual retrieved context. A third includes a controlled intervention result. A fourth includes a probe result with stated limitations. Observers can be asked which claims each packet supports. The aim is to test whether they distinguish levels of evidence, not to reward technical vocabulary regardless of its relevance.

Mechanistic evidence can also be overinterpreted philosophically. Identifying a computation associated with a response does not by itself settle whether the system has subjective experience. Different theories connect architecture and experience in different ways. The paper should not use an internal visualization as a direct picture of consciousness. It should state the theory-dependent inference if one is proposed and distinguish it from the measured effect.

At the same time, mechanistic limitations should not be used to dismiss the entire enterprise. A controlled intervention that reliably changes a specific behavior can be informative even if it does not explain every aspect of the system. The research program can accumulate such findings and refine its models. The observer study should therefore include cases where stronger internal evidence legitimately supports a stronger conclusion, not only cases designed to expose overreach.

The practical lesson is that access and interpretation are separate achievements. More access can reduce some forms of opacity while leaving conceptual and philosophical questions unresolved. Xenopsychology should study both the artificial system and the inferential practices through which internal evidence becomes an explanation. That is the appropriate response to a powerful image of an unfamiliar mind: not rejection of the image, but a clear account of what it represents and what it does not establish.

“Measurements need controls, interventions can have broad side effects, and a visually compelling pattern may not play the causal role the observer assigns to it.”

Rob Emerick · Solaris · §17

18. Consciousness claims require an explicit theoretical bridge

A system's behavior may prompt questions about consciousness, but the move from behavior to experience requires a theoretical bridge. The bridge might involve functional organization, particular computational properties, or other criteria. The paper does not select one theory as settled. It asks that any such inference identify its premises and the evidence relevant to them. Without that step, a statement about experience can appear to follow directly from a moving or surprising interaction when it does not.

Butlin and colleagues' report on AI consciousness illustrates an approach based on indicators drawn from scientific theories. It also makes clear that an indicator framework does not turn every matching property into conclusive proof. We cite it as a precedent for theory-explicit assessment, not as a current judgment about every model or as a result of this paper's proposed studies. [4]

The observer experiment should not pretend to score consciousness judgments against definitive hidden truth. It can examine whether participants distinguish the claim from more directly testable propositions, whether they identify theoretical assumptions, and whether their confidence responds to relevant rather than irrelevant evidence. A disclosure about a model's memory may matter to one continuity claim without resolving consciousness. A change in visual presentation may alter intuition without adding architectural evidence. Those distinctions are measurable at the level of reasoning about claims.

There are risks in both over-attribution and premature dismissal, but the paper does not assume they are equal in every context. Their importance depends on the decision, the evidence, and the theoretical commitments involved. A responsible analysis should state those dependencies. It should not use the existence of uncertainty to justify a preferred conclusion automatically. Uncertainty is a condition to reason under, not a conclusion in itself.

Ordinary social language can coexist with this caution. A user may say that an assistant seems thoughtful without intending a rigorous consciousness claim. The study should ask what the user actually infers and does. Research language, however, needs greater precision when it makes public claims about systems. The institution should avoid turning a metaphor into a finding merely because the metaphor is central to its brand.

This section preserves the deepest Solaris-inspired question while refusing to answer it through atmosphere. Some aspects of another mind may remain inaccessible to current methods. The appropriate response is to identify which theories connect available evidence to the claim, what would strengthen or weaken that connection, and which uncertainty remains. That is more intellectually demanding than either declaring a new consciousness or insisting that nothing interesting could be present.

19. Translation and documentation are part of the evidence chain

An encounter is rarely available without mediation. A reader encounters a novel through an edition and possibly a translation. An investigator encounters a model through an interface, a logging system, a selected transcript, and a description of the task. Those mediations do not make interpretation worthless, but they determine which evidence survives. A polished transcript can omit timing, retrieved records, failed tool calls, or an interface warning that substantially changes the explanation of an exchange. The unit under investigation is therefore not simply a response detached from its production history.

The limited literary basis of this paper should be understood in that way. Its close-reading anchor is the officially hosted English opening chapter, translated by Bill Johnston, supplemented by Lem's published commentary. It is not a scene-by-scene analysis of every English translation or film adaptation. The resulting restraint is methodological rather than cosmetic: an interpretation should not borrow details from several versions and present their composite as a single primary text. Source identification specifies which encounter a reader is being asked to consider. [1] [2]

For artificial systems, documentation can be modeled as a chain of transformations. Let an original event record contain the input, available context, retrieved material, tool proposals, tool outcomes, and output. A public account may select some fields, summarize others, and add an explanatory caption. Each transformation can preserve or change the apparent relation between evidence and claim. An assessment should keep the original record distinct from the edited exhibit and the author's interpretation. Otherwise, a conclusion can appear well supported because the evidence was silently rewritten to resemble it.

Consider a constructed exchange in which an assistant apparently initiates a correction without prompting. The complete record contains a hidden interface instruction requiring a correction after a checker reports an inconsistency. A condensed transcript removes both the checker and the instruction. The remaining dialogue invites an interpretation of spontaneous self-monitoring. That interpretation may be false for this particular event even if self-monitoring is an interesting possibility in other systems. Restoring the omitted fields changes the explanation without changing the visible wording of the response.

The proposed observer study can manipulate this evidence chain transparently. Participants receive either an explicitly incomplete excerpt, the complete event record, or a summary whose selection rules are disclosed. The experiment should not pretend that all three packets are informationally equivalent. It should ask whether observers notice missing causal context, request the omitted fields, and revise the scope of their claims when those fields become available. Such a study examines how documentation supports or frustrates understanding, not whether one group is inherently more credulous.

A practical publication rule follows. Every public behavioral example should specify whether it is complete, edited for length, reconstructed from logs, or invented for illustration. The present paper's examples are constructed. This distinction matters because a plausible illustration can teach a method without constituting evidence that a deployed system actually behaved that way. A field interested in unfamiliar minds must be especially careful not to create unfamiliarity through its own editorial omissions.

20. Build evidence packets that distinguish absence from non-disclosure

An evidence packet is the material an observer may inspect before making a judgment. It should be designed as deliberately as the model task itself. A missing tool log can mean that no tool was used, that the log was not collected, that it was withheld, or that a recording failure occurred. Those alternatives support different inferences. Treating every blank field as proof of absence would confound the observer study before the first participant responds.

We propose four explicit field states: observed and present, observed and absent, not observed, and withheld with a stated reason. A fifth state can mark disputed provenance when two records conflict. These states are bookkeeping categories, not a theory of cognition. Their purpose is to prevent the interface from supplying unwarranted certainty. A packet saying that memory retrieval was not recorded should not be visually identical to a packet documenting that retrieval was disabled and unavailable.

The packet should also define its sampling window. Ten selected successful exchanges are not equivalent to ten consecutive exchanges from a declared session. An observer who sees

only the selected successes cannot estimate a success rate for the underlying system. The appropriate answer may be that the examples demonstrate possible behavior but provide insufficient information about frequency. The coding scheme should reward that scope distinction instead of treating numerical reluctance as an error.

For a hypothetical continuity claim, a useful packet contains the exact memory condition, whether the model version changed, which records were available, and which outputs were selected for presentation. For a claim about conflict detection, it contains the instruction sources, their priority, and the available escalation actions. Different claims need different evidence. Providing every possible field can overwhelm participants while still omitting the one fact that discriminates the relevant explanations. Packet design should therefore follow a claim-to-evidence map.

Negative evidence requires particular care. A search that finds no record supports a conclusion only relative to the search coverage and record completeness. The study can include one packet with an exhaustive synthetic event log and another with an incomplete log. In both, the requested event is absent from the displayed material. The justified conclusions differ: the exhaustive record may establish that the event did not occur within the simulated system, while the incomplete record supports only non-observation. This contrast supplies a concrete test of evidential reasoning.

Finally, packet versions should be immutable during a study. Corrections require a new version, an explanation of what changed, and a record of which participants saw which version. That procedure prevents an evaluator from silently fixing an ambiguous example after observing responses. It also makes replication possible. The scientific contribution is not a beautifully narrated transcript; it is a recoverable relation between the evidence provided, the claims elicited, and the conditions under which those claims were considered warranted.

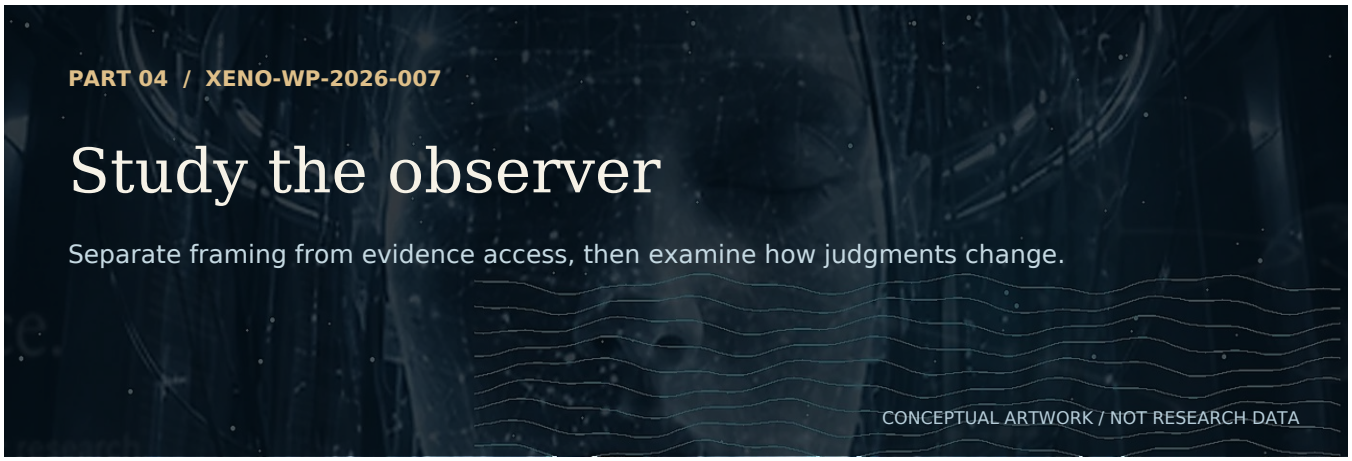
TABLE 2 / READING AID

Blank evidence fields are not interchangeable

Packet label	Meaning	Avoid inferring
Observed and present	The record documents the event.	That every possible cause is known.
Observed and absent	The observation supports absence within scope.	Absence beyond that scope.
Not observed	The relevant observation was not obtained.	That nothing happened.
Withheld, reason stated	Information was withheld.	That a blank means nonexistence.
Disputed provenance	Records conflict about the source.	A resolved causal story.

Bookkeeping states proposed by the manuscript. They describe evidence access, not the machine’s consciousness.

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



21. A proposed observer experiment with two separable manipulations

PROPOSED RESEARCH — NOT CONDUCTED

PROPOSED OBSERVER STUDY · NOT CONDUCTED

Framing and evidence access are separate manipulations

The central empirical proposal is a factorial observer study. Its first manipulation concerns framing: an otherwise identical artificial system is described in anthropomorphic, technical, or minimally descriptive language. Its second manipulation concerns evidence access: observers receive an output-only packet, a contextual record, or a contextual record plus a discriminating intervention. The design separates the influence of presentation from the influence of information. Neither manipulation is itself evidence about machine consciousness.

The same underlying synthetic episodes should appear across conditions, but an individual participant should not see several differently framed copies of the same episode without a design that explicitly addresses carryover. One option assigns different episodes to different conditions within participants while counterbalancing assignments across participants. Another assigns framing between participants and evidence access within participants. The choice affects what can be estimated and should be made before data collection, with a rationale tied to fatigue, learning, and contamination.

Manuscript passage · §21 · wording unchanged

Participants would first read an episode and produce a short account of what happened. They would then classify a set of narrowly worded propositions, identify evidence for their judgments, and choose what additional information would be most useful. After receiving a new packet, they would revise those judgments. The sequence measures not only initial attribution but also responsiveness to evidence. A participant who begins with an incorrect operational explanation and revises it appropriately should be distinguished from one who retains it after contradictory information.

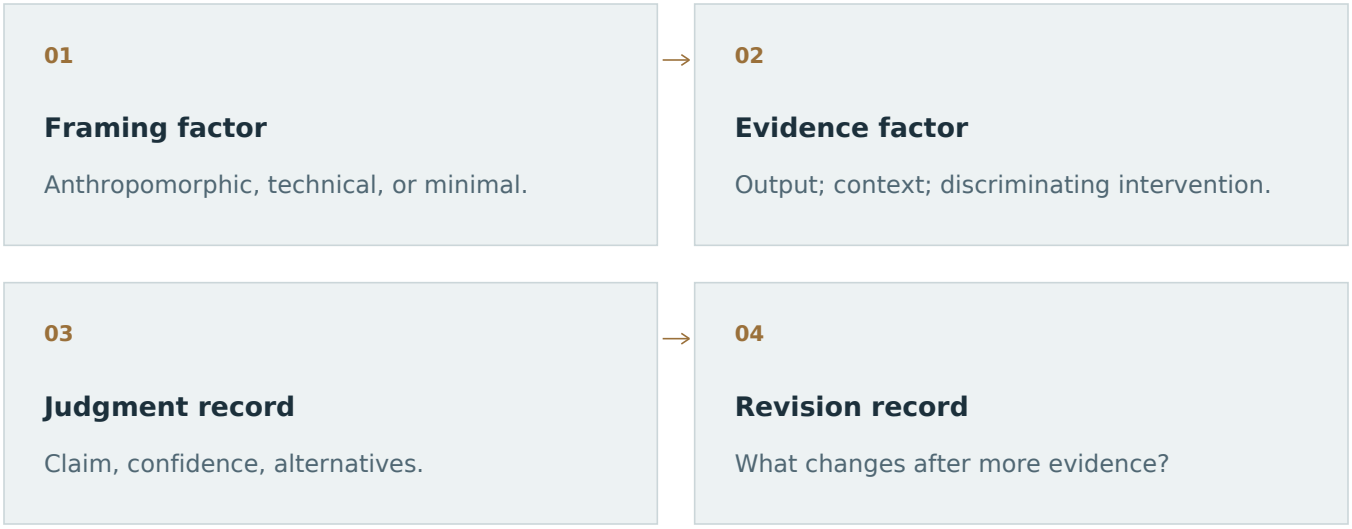
The propositions must vary in epistemic status. Some concern directly recorded events, such as whether a lookup tool returned a named record. Others concern a causal explanation supported by an intervention. A third group concerns broader interpretations for which the packet is insufficient. The study should not assign all claims the same kind of answer key. Recorded-event items can have a ground truth within the synthetic environment; theory-dependent mental-state claims require a different assessment of argument and evidential scope.

No recruitment has been conducted and no ethics approval is claimed. Before involving people, the researchers would need an appropriate review of consent, any incomplete disclosure, data retention, compensation, withdrawal, and debriefing. Framing an artificial system as a companion could affect expectations in ways that merit attention even in a short study. The design should avoid encouraging real emotional dependence, soliciting intimate disclosures, or implying that participants are interacting with a vulnerable conscious entity.

A pilot would assess comprehension of the packets and the burden of the task, not establish the main hypotheses. Pilot-driven changes to wording, exclusion rules, and coding would be documented before a confirmatory study. The number of participants should follow a defensible precision or power analysis informed by that design, not an impressive round number inserted into this paper. The proposed study is a route toward evidence. Its value here lies in specifying what an eventual result could and could not mean.

FIGURE 2 / CONCEPTUAL SCHEMATIC

A two-factor observer study



A schematic of the unexecuted observer study. Conditions require counterbalancing; the diagram is not participant data.

Reading aid based on §21, §22, §23. Source paragraphs remain in the full manuscript.

“The design separates the influence of presentation from the influence of information.”

Rob Emerick · Solaris · §21

22. An evidence-sensitive coding rubric without an ideological answer key

The coding rubric should evaluate the relation between a claim and its support rather than reward a favored philosophy of AI. A statement that a system has feelings is not automatically correct because it is sympathetic, and a statement that it is merely a lookup device is not automatically correct because it sounds skeptical. Either may exceed the available evidence. The coding task begins by identifying the proposition actually asserted and the kind of warrant offered for it.

We propose several independently scored dimensions. Event accuracy concerns whether a statement agrees with the complete synthetic record. Source accuracy concerns whether the observer identifies where the information came from. Scope calibration concerns whether a conclusion is restricted to the tested setting. Alternative sensitivity concerns whether the observer recognizes materially different explanations that remain possible. Revision quality concerns whether new evidence changes the relevant claim without causing unrelated beliefs to be discarded indiscriminately.

These dimensions should not immediately be collapsed into a single epistemic virtue score. An observer may be accurate about events but poor at distinguishing causal explanations. Another may correctly identify uncertainty while overlooking a decisive record. Reporting a profile preserves those differences. An aggregate score would require a justified use, a weighting scheme, and an examination of whether the dimensions can compensate for one another in that use. The rubric is proposed, not a validated psychological instrument.

Borderline cases require explicit examples. If an observer says the assistant remembered the rule, the coder should examine whether the surrounding text uses remembered as shorthand for successful retrieval or as a claim about human-like recollection. The same word can function at different levels. A follow-up question may be more informative than assigning an anthropomorphism label from vocabulary alone. Shanahan's discussion of language about large language models helps motivate this sensitivity to how ordinary descriptions can invite stronger inferences than their evidence warrants. [3]

At least a portion of responses should be independently coded, with disagreements retained and examined rather than resolved invisibly. The coding guide should include positive and negative examples created before the main analysis. Coders should not see participant condition when practical. Blinding cannot remove every interpretive judgment, but it can reduce opportunities for the expected effect to shape classification. A report should state where blinding was impossible and why.

The rubric's own limits are part of the research question. If coders cannot reliably distinguish functional shorthand from an experiential claim, the study should not publish a precise prevalence estimate as though the distinction were unproblematic. It may need revised elicitation questions or a narrower claim category. A science of interpretation should expose its interpretive instruments to the same scrutiny it applies to its subjects. Otherwise, the observer study merely relocates the Solaris problem from participant to researcher.

23. Statistical units, uncertainty, and the danger of counting sentences as people

The study contains several nested units: participants, underlying episodes, evidence packets, claims, and repeated judgments. These units are not interchangeable. A hundred coded sentences from one participant do not provide the same information about a population as a hundred independent participants. Likewise, many paraphrases of one synthetic episode may share the same ambiguity. An analysis that treats every sentence as independent would exaggerate how much the study has learned.

The primary estimands should be defined at the level that matches the question. A framing effect might concern the difference in a specified attribution rate across assigned descriptions, averaged over the sampled participants and episode families. An evidence-access effect might concern the probability of selecting a causally informative next test. A revision effect might concern changes in confidence for propositions whose support was altered by the new packet. Each estimand needs a declared denominator and a defined population of episodes.

A preregistered analysis could use a model that accounts for participant and episode variation, supplemented by transparent condition-level summaries. The paper does not prescribe a single statistical family before the response format and sampling plan are settled. Binary decisions, ordinal confidence ratings, and open-ended argument codes require different treatment. Whatever model is chosen, the report should show the raw pattern clearly enough that readers can see whether a headline effect depends on one unusual episode.

Missing responses deserve their own account. A participant may leave a question blank because the packet is difficult, because the response is optional, because of a technical failure, or because they reject its premise. Those reasons should not all be recoded as incorrect. The instrument should offer an explicit insufficient-evidence option when appropriate and record technical failures separately. Exclusion rules must not be adjusted after looking at which condition benefits from them.

The planned comparison set should also be limited. A study that examines many labels, outcomes, episode classes, confidence thresholds, and subgroups can produce an apparently striking pattern by selection alone. The distinction between confirmatory and exploratory analyses should be visible. Exploratory findings can guide later work without being advertised as if the exact hypothesis had been specified in advance. The paper's proposed design therefore includes an analysis inventory, not only a list of interesting questions.

Precision is especially important when drawing conclusions about observers' reasoning. A small or uncertain framing difference would not prove that presentation is irrelevant. A large difference in one task would not establish a universal tendency across all human–AI interaction. Intervals, episode coverage, and alternative specifications help state the result at the right scale. This methodological caution is consistent with the broader need for reliable and valid instruments in machine-psychology research, although the observer study measures people interpreting AI rather than administering a human test directly to a model. [7]

24. Competing interpretations should be allowed to win

A fair experiment must make it possible for evidence to support a richer explanation as well as a simpler one. If every episode is secretly a lookup table and every anthropomorphic framing is misleading, the study can demonstrate one kind of over-attribution but not the general quality of observers' reasoning. Its construction would predetermine the rhetorical lesson. A balanced stimulus set should therefore contain cases in which additional evidence genuinely strengthens a functional claim and cases in which it weakens one.

For example, an output-only packet may be compatible with either a stored answer or context-sensitive integration. A later packet can document a controlled manipulation that changes the response in a way predicted by the integration account. The observer should be permitted to increase confidence in that bounded causal explanation. This does not establish subjective experience. It does establish that skepticism should respond to discriminating evidence rather than remain a fixed posture.

Conversely, a fluent explanation may attribute an answer to a principle that the actual intervention does not support. The appropriate response is to question that explanation, not necessarily to deny all competence displayed in the task. Turpin and colleagues provide a relevant empirical warning that generated chain-of-thought explanations can omit influences on answers in tested settings. The current proposal uses that warning to separate explanatory self-report from independent evidence, without assuming that every explanation from every system is false. [5]

The framing conditions should not encode obvious caricatures. An anthropomorphic description that explicitly claims feelings, contrasted with a technical description that accurately discloses the architecture, changes both style and factual content. A cleaner comparison keeps factual disclosure constant while varying relational language or presentation. Where factual content cannot be matched, the difference must be acknowledged as part of the treatment rather than attributed solely to anthropomorphism.

Normative choices remain even in a careful design. The consequences of over-trusting a tool proposal differ from the consequences of an unwarranted claim about moral status. The study should identify which practical decision its confidence judgments inform. It should not assume that one universal threshold serves every context. A participant may reasonably demand different evidence before authorizing an action than before entertaining a speculative hypothesis in a seminar.

The resulting research would be less rhetorically tidy than a demonstration that people are simply fooled by machines. It would also be more useful. Xenopsychology should be able to discover that some observers revise appropriately, that some technical framings encourage unjustified dismissal, and that the important effect depends on the claim being judged. An institution devoted to understanding unfamiliar minds should not make its own favored interpretation immune to the encounter.

25. AI-assisted interpretation creates a second object of study

An investigator may ask another language model to summarize transcripts, identify contradictions, or propose explanations. That assistance can make a large record easier to navigate. It also introduces a second artificial system whose selections and interpretations require evaluation. The output of the assisting model should not be treated as an independent observation merely because it is written in an analytical tone. It is another transformation of the evidence packet.

A useful division of labor begins with mechanically checkable extraction. The assistant can identify candidate passages, link them to exact record identifiers, and distinguish quotations from paraphrases. A human or independent program can then verify whether the cited passage exists and whether the summary preserves its qualifiers. More ambitious tasks, such as attributing motives or classifying consciousness claims, require greater scrutiny because their answer keys are less direct. Automation should follow the verifiability of the task, not its rhetorical importance.

The study could compare unaided observers with observers using a retrieval tool or a language-model assistant, while holding access to the underlying record constant. Relevant outcomes include factual accuracy, time, source checking, and the ability to identify unsupported inferences. Such a comparison would measure the performance of a human–tool arrangement. It should not describe an improvement as a change in the human participant's underlying intelligence or as proof that the assistant understands the record in a human way.

There is a particular risk of correlated interpretation. If an assisting model summarizes an episode and another model judges the participant's response, both may share habits of phrasing or explanation. Agreement can then reflect a common bias rather than independent validation. The proposed design should retain human adjudication for ambiguous claims and use independently constructed event keys where possible. Model-based coding can be reported as exploratory or auxiliary until its errors have been characterized against those alternatives.

A source-grounded assistant should also preserve uncertainty in its output. If the original record says a tool result is unavailable, the summary should not transform that into a confident statement that the tool failed. If several hypotheses remain possible, it should not collapse them into a single narrative for readability. These are testable documentation behaviors. They connect this paper's observer problem to the broader Xenopsychology concern with memory, representation, and communication across systems.

The reflexive point is important. An AI can help investigate AI without becoming a privileged interpreter of its kind. Nor does human interpretation receive automatic authority simply by being human. Both require evidence, transparent transformations, and opportunities for correction. The research program should compare interpretive arrangements under explicit conditions rather than assign epistemic superiority by origin. That stance preserves the field's openness while making its methods answerable to observable performance.

26. Reporting incidents without turning hypotheses into biographies

When a deployed assistant behaves unexpectedly, an organization often needs an explanation quickly. A vivid account of what the system wanted can be easier to circulate than a careful account of incomplete logs and competing causes. The resulting narrative may then influence engineering changes, user expectations, and future investigations. The practical problem is not only whether the initial explanation is wrong; it is whether the organization can still distinguish an observation from the story built around it.

We propose an incident report with four separate layers. The event layer records what occurred and what remains unobserved. The causal layer lists candidate explanations and the evidence for each. The intervention layer records tests or changes and their outcomes. The decision layer explains what action is justified despite remaining uncertainty. A conclusion should be able to move between these layers only through an explicit evidential step. The format is a proposed reporting practice, not an existing certification standard.

Consider an invented incident in which an assistant tells two users incompatible versions of a project's status. A biography-like explanation says that it tried to please everyone. That

may be a hypothesis about a recurring behavioral tendency, but the event also admits more specific alternatives: different retrieved records, stale state, an ambiguous status field, or differently scoped questions. The report should identify which alternatives can be tested and what a successful test would change. A familiar human motive should not end the investigation prematurely.

An engineering response can be justified before every deeper question is settled. If a stale record demonstrably contributed, the organization can improve state handling while continuing to study the wider behavior. If several causes remain possible, it can restrict the affected action or require independent confirmation. Such decisions should be described in terms of the evidence and the task, not as proof of a permanent personality trait. The intervention may improve the system even when the philosophical interpretation remains open.

The report should also distinguish remediation from retrospective explanation. A new instruction may prevent recurrence without showing why the original event occurred. Conversely, a convincing reconstruction may reveal a cause that is difficult to remove. Conflating these achievements can lead to overconfidence. The proposed incident format records both: what the change demonstrably does, and which explanation it supports. A test that only shows improved outcomes should not be advertised as having revealed the system's motive.

This operational use gives the Solaris argument a concrete destination. The limits of understanding are not an excuse for theatrical mystery or paralysis. They are a reason to preserve evidence, qualify explanations, and design actions that remain sensible under several live hypotheses. An organization can become better at working with artificial systems without inventing an inner biography for every failure or pretending that implementation details make interpretation unnecessary.

CONCEPT IN THIS PAPER

Interpretive Overreach

Keep the observed incident, possible mechanisms, and next tests apart. A compelling explanation is not a completed evidence chain.

Reading aid based on §9, §26. Source paragraphs remain in the full manuscript.

Open Lexicon entry

27. Objections, boundary conditions, and ways the program could fail

OBJECTIONS & LIMITS

What the observer study could fail to show

A first objection is that the proposed observer study measures familiar reasoning biases rather than anything distinctive about artificial intelligence. That is a serious possibility. The design should include non-AI comparison cases where appropriate and explain which features of AI presentation, opacity, or interaction are hypothesized to matter. If the same pattern appears across every kind of unfamiliar system, the contribution may be a general account of evidence interpretation rather than an AI-specific effect. That would still be informative, but it should change the claim.

A second objection is that synthetic episodes are too clean. Their controlled records and known event keys may not resemble the ambiguity of deployed systems. The response is not to abandon synthetic control immediately. It is to state the division of work: controlled episodes test whether observers can use specified evidence, while later naturalistic studies examine whether the relevant evidence is available and usable in practice. Success in the first setting is not proof of transfer to the second.

Manuscript passage · §27 · wording unchanged

A third objection concerns the coding rubric. Researchers may classify as overreach the interpretations they personally dislike. Independent coding, transparent examples, proposition-specific keys, and explicit theory dependence reduce this risk but do not eliminate it. Some judgments should remain contested rather than forced into a consensus score. A strong report would publish representative disagreements and show how conclusions change under reasonable alternative codings. The study's own interpretation must remain reviewable.

A fourth objection is that asking participants to explain evidence changes the phenomenon. Ordinary users do not constantly construct claim ledgers during conversation. The proposed task therefore measures reflective interpretation under elicitation, not every spontaneous reaction to an assistant. A later study could compare immediate impressions with judgments after structured review. The difference might show where a documentation interface helps, but it should not be described as a direct map of unprompted social experience.

A fifth objection is that philosophical neutrality conceals practical commitments. The paper does have commitments: evidence should constrain claims; different claim types need different warrants; and uncertainty should be made visible. It does not claim that all theories of mind are equally plausible or that every ethical response to uncertainty is equivalent. It deliberately leaves those broader debates open while specifying a narrower empirical program. Readers should be able to accept the proposed experiment without accepting every part of the institution's vocabulary.

The program could fail constructively. Framing might have little effect once information is matched. Participants might distinguish evidence levels more effectively than expected. The rubric might prove unreliable. Synthetic controls might reveal that purported attribution errors are actually reasonable interpretations of ambiguous wording. Each outcome would improve the research question if it were reported honestly. A working paper earns its place not by guaranteeing a dramatic finding but by exposing its claims to outcomes that could require revision.

28. Conclusion: understanding includes knowing what the evidence cannot decide

Solaris offers this collection a final reversal of perspective. The unfamiliar system is not the only source of difficulty; the observer arrives with categories, narratives, instruments, and desires for explanation. The literary work does not supply empirical evidence about present AI. It provides a disciplined reason to ask how interpretation proceeds when the object of inquiry resists familiar accounts. The research proposed here translates that question into evidence packets, competing explanations, and observable revisions of judgment.

The paper's principal contribution is a separation of several problems too often compressed into the word mystery. Missing observations, inaccessible interfaces, unknown mechanisms, inadequate concepts, and unresolved theories of experience require different responses. Some can be addressed by collecting a record. Others need an intervention or a better instrument. Some remain philosophical questions that cannot be scored against a hidden experimental answer key. Recognizing these differences is an achievement of understanding, not an admission that inquiry has failed.

The proposed observer experiment examines how labels and evidence access affect what people claim about artificial systems. Its outcome measures concern event accuracy, source attribution, scope, alternatives, and revision. It does not diagnose participants, establish machine consciousness, or validate a universal cognitive-distance scale. No participants have been recruited and no results are reported. The constructed cases demonstrate how

the design could discriminate explanations; they do not establish how often those explanations occur in current deployments.

This approach connects the seven papers without making them interchangeable. Darmok asks what shared context makes interpretation possible. Arrival examines assumptions in representation and task formulation. Close Encounters separates a channel from meaningful coordination and repair. Data distinguishes kinds of continuity and recognition. HAL separates directive conflict from the authority that turns an error into an action. The foundational paper locates these questions in the complete system. Solaris asks whether the observer's resulting explanation is warranted by the evidence actually obtained.

The institution's language should evolve through that discipline. Cognitive distance, synthetic personality, and cognitive translation are useful only when they direct attention toward distinctions that can be investigated or argued clearly. They should not become decorative substitutes for explanation. A successful research program would sometimes narrow its claims, replace its terminology, or discover that a supposedly novel problem is better understood through an established neighboring method. Such revision would be evidence of intellectual progress rather than a loss of identity.

Understanding an artificial mind need not mean making it human, accepting its self-description, or reducing every question to a single mechanical label. It can mean constructing a reliable map of what the system does, which conditions matter, which explanations survive comparison, and where uncertainty remains. That map will be partial. Its value lies in being explicit enough to test, useful enough to guide interaction, and open enough to change when a new encounter supplies evidence the previous account could not accommodate.

Open questions

1. Is the uncertainty due to missing observations, inaccessible mechanisms, inadequate concepts, or something else?

2. What would make an observer revise a psychological explanation of an identical record?

3. Which claim about the system remains unanswered even after a framing effect is measured?

References & source scope

Literary scope: the officially hosted opening chapter of Solaris, translated by Bill Johnston, and Lem’s cited author commentary. This is not a full-novel or cross-adaptation analysis. All observer studies are proposals; no participants have been recruited.

Primary creative text / excerpt

[1] **Stanisław Lem. Solaris — “The Newcomer,” translated by Bill Johnston. Official excerpt.**

The publicly hosted opening chapter is the close-reading scope, not the full novel. Translation credit is explicit on the source page. No scene-by-scene comparison of adaptations is claimed.

https://english.lem.pl/works/novels/solaris/47-look-inside-solaris

Author commentary

[2] **Stanisław Lem. Solaris — Lem’s opinion. Official author commentary.**

Author commentary provides a limited interpretive context; it does not exhaust the novel’s meanings or substitute for the full text.

https://english.lem.pl/works/novels/solaris/44-lems-opinion

Research

[3] **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

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

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

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

[5] **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] **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/

[7] **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/

Publication record

Rob Emerick (2026). Solaris — The Limits of Understanding a Truly Other Mind. XENO-WP-2026-007, 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/solaris-limits-of-understanding-other-mind

Full paper: https://xenopsychology.com/insights/solaris-limits-of-understanding-other-mind/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