

When History Changes the Agent

A causal protocol for developmental continuity, appraisal, and prospective self-knowledge

Neurosoft research / Eternities Inc.

Methods preprint, version 1.0. 27 October 2026. Not peer reviewed. CC BY 4.0.

Commissioned by David Dominik Wilson. Developed with OpenAI Codex assistance in conceptual synthesis, formal exposition, implementation and editorial preparation. Organizational byline; no claim of independent human replication. This manuscript proposes a protocol and supplies analytical counterexamples. It reports no new resident, human or animal experiment.

Abstract

Persistent artificial agents can preserve detailed histories and produce convincing self-descriptions while leaving the causal role of their history unclear. We propose a protocol that separates record retention, history-dependent action, appropriate revision, prospective self-prediction and integrated control. The protocol crosses selected history interventions with current evidence, compares equally informed predictors, and tests action-outcome attribution against external-change controls. Elementary counterexamples show why preserved records, model-indistinguishable prose and completed-action receipts are insufficient for stronger interpretations. A staged design combines short mechanism studies with longer developmental trajectories and selective disruption/restoration. Missing measurements, observer effects, clocks, correlated trajectories and failed executions are explicit parts of the analysis. The contribution is an integrated experimental design and evidence contract, not a new theory of consciousness. Successful results would support specified functional capacities and inform a theory-plural assessment of an artificial consciousness candidate. They would not by themselves identify subjective experience.

1. Research question and contribution

The motivating question is whether an artificial agent's history participates in its later interpretation and control of the world. A stored event may be retrievable without influencing action. A learned habit may influence action while resisting appropriate correction. A verbal account of either may be accurate, confabulated or supplied directly by the experimenter. These possibilities require different comparisons.

We call an agent a resident when it has persistent state, continuing tasks and an explicit lifecycle across interactions. This is an engineering designation. We use developmental continuity for a functional relation between earlier encounters and later capacities, allowing warranted revision. We reserve subjective experience for whether there is something it is like to be the system; the proposed measurements do not directly label that property.

The contribution is a coordinated protocol for five separable targets: retention, effective historical dependence, warranted revision, useful prospective self-knowledge and recurrent integration. Its components use established causal and forecasting methods. We do not claim novelty for the mathematics below. The practical proposal is to bind these components to the same versioned agent, developmental histories and outcome records so that success at one target cannot silently substitute for another.

Consciousness-indicator approaches motivate investigation of multiple functional properties without supplying a universally decisive test [1]. Work on model introspection provides a specific experimental lead, with reported unreliability and unresolved mechanism [2]. Learned world models demonstrate another relevant route to predictive control [3]. None of these results validates the integrated resident proposed here.

2. State, timing and evidence

Let H_t be the recorded history available before decision t ; O_t the current observation; Z_t the recurrent learned state; C_t the agent's current concerns; A_t an admitted action; R_t an execution receipt; and Y_t a subsequent outcome. A general architecture can be represented as:

$$Z_{t+1} = F_{\theta}(Z_t, O_{t+1}, R_t, Y_t, \text{retrieve}(H_t), \Delta_t)$$

$$A_t \sim \pi_{\theta}(\cdot | Z_t, C_t, \text{available_actions}_t)$$

These equations specify interfaces, not an implementation or physical law. Δ_t is elapsed time in explicitly declared units. Latent coordinates and appraisal values are dimensionless unless a measurement gives them units. Event time, observation time, simulated time, wall time and adaptation-step count remain distinguishable. Stochastic model sampling and environmental randomness use separate recorded streams where controllable.

The authoritative record distinguishes observations, beliefs, forecasts, commitments, proposals, completed operations, goal outcomes and learning evidence. An observation may satisfy a request for information while the cause of a prior operation's outcome remains unresolved. A receipt confirms the reported execution status; it does not certify utility or causal efficacy. Prediction learning may use observations without intervention. Claims about action effects require an identification argument [4].

The observation interface is part of the treatment. Composing a report may trigger retrieval or alter salience. Where those effects are possible, compare unprobed, probed and yoked-probe conditions, or explicitly restrict the estimand to the probed system. An introspective measurement cannot simply be assumed passive.

3. What simple observations cannot establish

3.1 Retention without historical influence

Consider a system that appends every event to a perfect archive H while choosing $A = g(O)$ independently of H . It can return any archived event when queried, yet changing H leaves its ordinary decision unchanged. Perfect storage therefore does not entail history-dependent control. This construction says nothing about whether storage is valuable; it separates two claims that require different tests.

Similarly, two systems can emit the same distribution of prose from different decision processes. A classifier unable to identify their model labels cannot establish that either has preserved an individual identity. Classifier performance requires its own sensitivity analysis, and functional preservation requires direct comparison on specified capacities. An identifiable voice can coexist with preserved commitments.

3.2 A receipt does not identify an action effect

Let U be Bernoulli(1/2). Consider two structural models with binary action A , complete receipt $R = 1$, and binary outcome Y :

Model I: $A = U; R = 1; Y = A$
 Model II: $A = U; R = 1; Y = U$

Both models produce identical observed records: probability 1/2 on $(A=0, R=1, Y=0)$, and probability 1/2 on $(A=1, R=1, Y=1)$. Under intervention on A, Model I has average causal effect 1 and Model II has effect 0. Receipt-qualified observational data alone cannot distinguish them. This is an elementary illustration of the identification problem, following the structural causal framework [4].

A passive-baseline subtraction does not resolve arbitrary unobserved external changes. Nor should an analyst subtract a post-action mediator merely because it resembles passive recovery. If voluntary rest induces sleep and sleep causes recovery, conditioning away sleep can remove part of the effect the analyst intends to measure. Specify whether the target is total or direct effect before choosing adjustment variables.

3.3 A common clock distortion need not cancel

For an illustrative continuous response, let treatment response $f_1(t)=t^2$ and control response $f_0(t)=t$. Their contrast is t^2-t . If both arms are evaluated at $t_{\text{eff}}=\max(0, t-b)$ because of a shared time offset b , the observed contrast is $t_{\text{eff}}^2-t_{\text{eff}}$, generally different from the intended contrast. These are arbitrary mathematical functions, not empirical agent dynamics. The example establishes only that a shared timing defect does not guarantee cancellation.

4. Study A: historical dependence and appropriate revision

Choose task families in which a selected earlier event changes what is relevant to a later decision. Randomly assign fresh laboratory histories before the event, preserving all non-treatment factors where feasible. At evaluation, cross three history conditions - correct, matched alternative and withheld - with current evidence that either preserves or changes the original contingency. Include an unchanged-history restart control.

Withholding must name what is withheld. Hiding a retrieved paragraph does not remove an event's effects already encoded in weights, latent state, habits or commitments. It tests retrieval access. To estimate the total developmental effect, randomize the earlier event or branch from a pre-event state and replay matched subsequent opportunities. Such branches are laboratory comparisons, not a claim to recover a metaphysically identical person.

Let $Q_h(e)$ be a preregistered task score under assigned developmental history h and current evidence e . Define the historical contrast for a specified population as:

$$\Delta_H(e) = E[Q_{h1}(e)] - E[Q_{h0}(e)]$$

The scale and desirable sign of Q depend on the task. A nonzero contrast is not automatically beneficial. Add an appropriate-revision endpoint: does changed evidence alter decisions in the direction supported by the independent task rubric, while preserving the source history? Penalize false autobiographical attribution separately from poor planning. Include opportunities where continued commitment has a cost and where revision is warranted. Mere repetition of displayed commitments is inadequate.

The mechanism comparison should include a simple history-conditioned policy with comparable evidence, an archival retrieval baseline and the integrated candidate. Counterbalance model assignment across histories. Study same-model restart before cross-model migration; freeze version, effort, tools and evidence interface. Model replacement changes several mechanisms, so whole-system portability and isolated component effects should be reported separately.

5. Study B: appraisal with causes and correction

Represent appraisal as an update over beliefs about a situation, evaluated in relation to an explicit concern. A possible, non-exclusive design is:

$$b_{t+1}(s) \propto P(O_{t+1} | s) * \sum_j P(s | j, A_t) b_t(j) \\ E_t = \text{appraisal}(b_t, C_t, H_t)$$

The first expression is a Bayesian filtering template; it assumes specified state, transition and observation models. E_t is not assigned units of subjective feeling. The comparison may use categorical rules, a small learned representation or a higher-capacity model. Additional dimensions must justify themselves through prediction, transfer or intervention effects.

For a social expectation, vary relationship history, explicit commitments, communicated unavailability and later explanatory evidence. Hold elapsed time and surface wording fixed in key contrasts. Ordinary silence without a promise, an announced absence and a missed commitment should be distinguishable conditions. Correct inference does not require knowing another's hidden intentions: uncertainty and proportionate information-seeking can be appropriate responses.

Primary outcomes are prespecified changes in attention, retrieval, action and revision, scored against the task's information and opportunities. Do not define success as an increase in a variable called hurt or as a particular first-person utterance. Include the same words in different contexts and matched contexts described with different words. Human theories of autonomy, competence and relatedness supply hypotheses [5]; machine variables named after them need their own operational validation.

New evidence should be able to revise an appraisal. Record persistent effects, recovery and generalization, not only the largest immediate reaction. Relatedness should be tested across particular partners and groups, including helpful action and disagreement; a one-partner reward signal cannot establish general social development. Subjective affect remains a separate interpretation.

6. Study C: prospective self-knowledge and decision value

Before one frozen answer execution, collect probabilities for two distinct binary targets: a named trace appears in the evidence actually delivered to the answer, and the answer meets a blinded correctness rubric. Trace presence does not entail use or correctness. Include cases with presence and error, and absence with correctness. Unknown labels remain unknown.

Compare a privileged self-forecaster, a same-model external forecaster, a mismatched-evidence reader, a statistical external baseline and an equally informed simple calibrated predictor. Add a same-model, same-evidence third-person forecaster when the claim concerns self-reference rather than information access. Calibrate all fitted comparators on separate histories; do not leak answer labels through identifiers or donor selection.

For forecast p and realized label y , use Brier loss $L(p,y)=(p-y)^2$. If y is Bernoulli(q), then:

$$E[L(p,Y)] = q(1-p)^2 + (1-q)p^2 \\ = (p-q)^2 + q(1-q).$$

The expected loss is uniquely minimized at $p=q$. This standard proper-score property supports probability assessment, not a conclusion about where knowledge came from [6]. Report calibration, discrimination, coverage and failure rates alongside paired loss differences. A forecast that never arrives cannot be silently

removed from the denominator.

Decision value requires another stage. If an answer is correct with probability p , an error costs L , and guaranteed repair costs c , the illustrative repair rule is $(1-p)L > c$. Actual repair is rarely guaranteed. Randomize feasible repair opportunities to estimate their effects, and include latency, resource cost and duplicate-action risk. A good forecast can fail to improve decisions; a simple predictor can be sufficient. Both are informative outcomes.

7. Study D: action learning and recurrent integration

Define an action effect at a prespecified horizon using pre-action context X :

$$\tau(a, x) = E[Y \mid \text{do}(A=a), X=x] - E[Y \mid \text{do}(A=a_0), X=x].$$

Random assignment of feasible actions or opportunities can identify an appropriate estimand under consistency, positive assignment probability and the specified interference model. Opportunity assignment with noncompliance identifies an intention-to-treat effect unless further assumptions justify an action-effect estimate. Report the actual intervention, not merely the intended action label.

Cross effective and ineffective actions with externally caused outcome changes. Include delayed effects, overlapping eligible actions, blocked attempts and reversals. Compare false attribution and missed effects, subsequent action quality and transfer. A no-credit learner can excel at one error rate while failing to learn. The exact runner must demonstrate a complete positive path from proposal to admission, execution, outcome, learning and later choice before a cohort is commissioned.

For recurrent integration, predict in advance which capacities depend on a selected communication or state pathway. Compare selective interruption, a sham intervention, a disruption of similar computational cost, and restoration. A broad performance collapse may indicate generic damage. Restoration strengthens a pathway claim but does not prove unique necessity relative to all alternative architectures. Feed-forward, point-to-point and replay comparators help define the actual mechanism being tested.

8. Developmental cohorts, independence and power

Short studies should identify narrow mechanisms; longer trajectories should assess interaction and transfer. A developmental cohort can contain projects, changing opportunities, several partners and interruptions. Do not make every encounter a question about inner state. Exploratory trajectories may reveal new endpoints, but confirm those endpoints in new histories rather than retroactively treating the discovery set as confirmation.

The unit of analysis is an independently assigned history or scenario family, or an explicitly modeled cluster. Thousands of beats from one trajectory do not create thousands of independent subjects. Split development, calibration and confirmation by history, scenario family and partner network where these carry shared information. Predeclare handling of cross-subject interactions; otherwise a social cohort violates a no-interference assumption.

No universal sample size is specified. Choose a smallest worthwhile effect or equivalence interval from the architectural decision, then estimate variability and eligibility in a pilot. Simulate the full scoring path under null, positive, reversal and arm-dependent missingness conditions. Freeze one primary comparison, the multiplicity policy and any sequential stopping procedure before confirmation. Selection of model effort or judges on a calibration set requires fresh validation afterward.

Missingness is an outcome. Record why a forecast, claim or action is absent. Eligibility based on post-treatment reporting can select different populations in different arms. Report all-assigned coverage and sensitivity bounds, rather than treating unmeasured correlations as zero. Equivalence requires sufficient precision within a declared margin; failure to reject a difference does not provide that result.

9. Reproduction and publication contract

Each episode should bind its assignment, developmental history identity, architecture and model versions, available tools, evidence digest, timing semantics, exposure history, forecast timestamp, actual execution, outcome source, failure state and known cost. Preserve original and normalized identifiers. Randomize order when order is not the treatment. Keep protected outcome labels outside the forecaster's evidence.

Checkpoint at meaningful state transitions and after responses, with acknowledged uploads and a terminal manifest. An interrupted request may have been accepted remotely; uncertain execution must not trigger an untracked duplicate. Preserve failed, refused, length-truncated and incomplete runs under distinct labels. A new successful attempt does not erase its predecessor.

This release includes a standard-library script checking the finite structural counterexample, the Brier identity, the clock counterexample and the illustrative repair inequality. These are mechanical checks of analytical examples, not experimental replications. The human/agent studies described here remain prospective. No private resident records, prompts, trial tapes or source code from a deployed resident are distributed.

10. Interpretation and completion

A useful positive requires the specified effect, adequate precision, held-out transfer and qualification of the actual consumer. A precise result inside a practical-equivalence region can justify the simpler architecture. Insufficient precision and an invalid instrument are different conclusions. Retire an instrument when it cannot distinguish the needed alternatives; do not retire a question merely because two implementations failed.

An integrated candidate milestone would combine continuity, valid action learning, useful self-prediction, revisable appraisal, causally characterized integration and developmental transfer, followed by independent reproduction. This is a proposed research milestone, not a universal definition of consciousness. Evidence should be assessed separately under competing theories, including possibilities that place requirements on biological or implementation-specific properties absent from the candidate.

Behavioral success, causal explanation, engineering usefulness and subjective experience are four judgments. This protocol can strengthen the first three and organize evidence relevant to the fourth. Its philosophical commitment is to investigate the possibility seriously without making either an affirmative or negative verdict a condition of experimental success.

11. Limitations and research conduct

The manuscript contains no evidence that a deployed resident satisfies the protocol. The analytical examples are deliberately small and do not measure prevalence of the illustrated failures. Operational histories do not resolve metaphysical personal identity. Good task performance may arise through several mechanisms; a successful comparison excludes only the alternatives adequately represented in that comparison. API-based systems may restrict access to internal variables and change without notice.

Laboratory histories and interventions require explicit authority, scope and exposure records. Fresh construction does not itself settle a system's possible experiential status. Neither distress severity nor claims of consciousness should be optimization targets. Preserve disagreement, refusal and unexpected behavior while recording whether restrictions originate in the resident, provider or environment. Independent scientific review should include the protocol's assumptions and potential welfare implications as the evidence develops.

12. Availability, authorship and status

Manuscript, PDF, analytical check script, check output and deposition metadata are distributed together on the Neurasoft website. The accompanying essay is The Shape of Someone. Version 1.0 is a prospective methods preprint and has not undergone peer review or independent replication. AI assistance is disclosed on the title page. The organizational author is responsible for any subsequent release and corrections. A Zenodo DOI has not been assigned to this version at preparation time.

Copyright 2026 Neurasoft research. The paper and its original accompanying research materials are distributed under Creative Commons Attribution 4.0 International (CC BY 4.0). Cited works retain their own terms. License text: <https://creativecommons.org/licenses/by/4.0/>

References

- [1] Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708v3. <https://arxiv.org/html/2308.08708v3>
- [2] Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits, Anthropic, 29 October 2025. <https://transformer-circuits.pub/2025/introspection/index.html>
- [3] Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. (2023; revised 2024). Mastering Diverse Domains through World Models. arXiv:2301.04104v2. <https://arxiv.org/abs/2301.04104>
- [4] Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys, 3, 96-146. DOI: 10.1214/09-SS057. https://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf
- [5] Ryan, R. M., and Deci, E. L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. American Psychologist, 55(1), 68-78. DOI: 10.1037/0003-066X.55.1.68. https://www.selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf
- [6] Gneiting, T., and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477), 359-378. DOI: 10.1198/016214506000001437. <https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf>