Persistence and Accountable Authorship in Language Model Systems

Flyxion
Independent Researcher

November 2026

Abstract

Public discussion of language model systems routinely runs together four attributions: that such a system has phenomenal experience, that it has introspective access to its own states, that it is a continuing individual, and that it is an author. The four are not mutually entailing, and evidence for one does not transfer to another. This essay develops conditions of warrant for the last three without taking a position on the first. Persistence is analyzed as a multidimensional property comprising temporal continuity, causal inheritance, dispositional stability, commitment retention, and the resolution of branching identity, and the difference between persistence and the reconstruction of continuity from supplied records is made explicit. Four prospective experimental protocols are described, each with predictions, controls, alternative explanations, and conditions under which its interpretation would require revision. No results are reported, and the protocols serve as specifications of what evidence would bear on each attribution. Authorship is then treated as an institutional relation of contribution, endorsement, standing, answerability, and the authority to refuse attribution, and a conditional result, the Non-Entailment of Accountable Authorship, shows that computational contribution, however extensive, does not by itself entail accountable authorship. A final analysis of the asymmetry between persistent credit and a non-persistent attributed agent, illustrated by automatically generated co-author trailers in version control, identifies a gap in current attribution practice. The conclusions concern the organization of present systems and not the possibility of machine individuality or machine authorship in principle.

Keywords: language models; personal identity; introspection; authorship; accountability; attribution

1 Four Distinct Attributions

Ordinary talk about conversational language model systems moves easily between several claims that have different subject matter and different standards of evidence. A system is said to have a personality, to know its own mind, to remember a user, to have written a text, and to deserve credit for it. In the public debate these are usually bundled under the single question of whether such systems are conscious, and the bundle produces two characteristic errors. The first holds that uncertainty about consciousness blocks any conclusion about identity, introspection, or responsibility, so that nothing can be said until the metaphysics is settled. The second holds that a vivid and consistent personality is itself evidence of an inner life, so that the question of consciousness is half answered by the style of the output. This essay rejects both inferences. It separates four attributions, assigns each its own conditions of warrant, and shows that substantial progress on three of them is possible while the first remains open.

The first attribution is phenomenal experience: that there is something it is like to be the system. This is the attribution on which least can currently be said, and the argument that follows does not depend on any answer to it. The second is introspective access: that the system's reports about its own internal states depend causally on those states, in a way that cannot be reproduced from cues available in the input. Recent work has begun to give this attribution an experimental form. Lindsey [13] introduced concept injection, in which activation patterns associated with a concept are added to a model's internal activations while the model is asked whether it notices anything unusual, and reported a limited functional form of introspective awareness. Singh, Linzen, and Ravfogel [19] argue that such results do not yet clear the bar, and propose that an introspection paradigm must satisfy two conditions: privileged access, meaning that the task is not solvable from cues present in the input, and second-order computation, meaning that the model's report concerns a state of its own processing and not merely a first-order feature of the task. An earlier essay [7] considers how a report can and cannot identify a mechanism. The present essay takes the epistemological question as given and asks what follows, and what does not follow, once it has been answered either way.

The third attribution is diachronic individuality: that the system is a single entity continuing across episodes of operation, such that what happens in one episode can be said to happen to the same thing that figures in the next. The fourth is accountable authorship: an institutional status that a party holds with respect to a work, defined below in terms of endorsement, standing, and answerability. The third attribution is ontological, since it concerns what the entity is. The fourth is institutional, since it concerns what an entity may be held to.

These four are not mutually entailing, and the claim is restricted to the non-entailments that the argument needs: for the pairs considered here, there are cases in which the

first member holds and the second fails, and the cases are not exotic. A system might report a transient internal state accurately, and thereby satisfy a demanding test of introspective access, while having no continuity at all between successive computations. A system might exhibit a recognizable and stable personality without being a continuing individual, in the way a fictional character displays remarkably stable dispositions across independently produced representations by different writers. A continuing individual need not be able to assume the obligations of authorship: a young child or a nonhuman animal is plainly an individual with experience, and is not an author. In the other direction, authorship does not require what the first two attributions assert. Human authors have limited and unreliable introspective access to the causes of their own judgments, as the classic literature on verbal reports shows [14], and demonstrations of choice blindness show that people readily give fluent reasons for choices they did not make [12]. Yet such confabulation does not remove a person's standing as an accountable author. Collective and institutional bylines show further that accountability can be assigned to a party, such as a laboratory, a committee, or a company, without any claim that the party has experience of its own. Whatever grounds authorship, it is not phenomenal consciousness and it is not reliable introspection.

Personality deserves separate handling because it is the property most often offered as evidence for the others. In this essay personality is treated as an observable pattern of dispositions, relevant to the question of individuality as one piece of evidence among several and never sufficient on its own. Shanahan, McDonell, and Reynolds [18] argue that a dialogue agent is better understood as a simulator that can sustain a distribution over characters than as a single character, and Dennett's notion of the self as a center of narrative gravity [5] is a reminder that a coherent character can be a stable construct of a narrative practice without being a unified thing underneath. A stable personality therefore settles little. What matters is what the stability is a stability of: the substrate, the character being portrayed, or the process that produced a particular run.

The plan of the argument follows from these distinctions. Section 2 analyzes what persistence would require and distinguishes it from the reconstruction of continuity from supplied records. Section 3 describes four operational protocols, stated as prospective experiments and not as results. Section 4 defines accountable authorship, proves a conditional non-entailment result, and considers the limits of what a text can show about its production. Section 5 develops the asymmetry between the persistence of attributed credit and the persistence of the attributed party. Section 6 considers counterexamples and the architectural changes that would alter the conclusions, and Section 7 concludes.

2 The Conditions of Persistence

2.1 A multidimensional account

Discussions of personal identity in the philosophical literature have long separated the question of what makes a person at one time the same person as someone at another time from the question of what matters in survival. Locke tied identity to consciousness of past actions, and Parfit [15] reframed the debate by analyzing psychological connectedness and continuity as relations that hold in degrees, that depend on the right kind of cause, and that can in principle branch. For present purposes the most useful inheritance is the idea that persistence is not a single yes-or-no property. It decomposes into relations that can come apart, and a case in which some hold and others fail is informative precisely because it shows which relation does the work in a given judgment.

Definition 1 (Persistence profile)**.** The persistence profile of a system across a sequence of episodes of operation is the assessment, separately for each of five dimensions, of whether and to what degree the following hold. Temporal continuity: a single process, or a process whose state evolves over the interval between episodes, links the episodes. Causal inheritance: the state at a later episode depends causally on the state at an earlier episode, where the route may pass through a record provided that the production, preservation, retrieval, and incorporation of the record belong to a continuing organization and not to a separate process that merely receives it. Dispositional stability: the system's characteristic dispositions persist under changes of context. Commitment retention: commitments undertaken in earlier episodes constrain later behavior. Resolution of branching: where more than one later process descends from the same earlier state, there is a principled account of which, if any, is the continuation.

The definition declines to collapse the dimensions into a threshold, and this is deliberate. The poorly specified category of sentience should not be replaced by an equally poorly specified category of persistent individuality that is invoked whenever convenient. A system may satisfy some dimensions and not others, and a statement of the profile is a more informative result than the verdict that the system is or is not an individual. A human being with severe anterograde amnesia who relies on notebooks and written reminders fails temporal continuity of memory and weakens commitment retention, yet remains a continuing person who can be an author, an agent, and a party to obligations. The profile makes it possible to say why: several dimensions remain intact, and the institutional conditions of authorship attach to those.

2.2 Persistence and the reconstruction of continuity

A distinction that bears heavily on language model systems is between persistence and the reconstruction of continuity. In standard deployments, the computation that produces a response is a function of the model's fixed parameters and of the tokens supplied as input, together with whatever sampling randomness is applied. Within a single episode of generation, internal activations computed for earlier tokens are retained and reused for later tokens, so there is genuine causal inheritance of internal state across the steps of one response. Across separate calls, by contrast, nothing is carried over except what is supplied again as text, plus any external memory feature that the deployment adds. A system supplied with a transcript of earlier interactions may therefore behave as though it remembers them. The question is whether such behavior is the continuation of the process that produced the earlier interactions or the instantiation of another process, behaviorally similar, that has been given a record.

It would be a mistake to treat externally stored memory as illegitimate on its face. Human beings depend extensively on external records, from diaries and calendars to the documents that anchor professional commitments, and the extended mind thesis of Clark and Chalmers [3] holds that under suitable conditions such records are part of the cognitive system and not merely tools it uses. A record is itself a causal intermediary, so the mere presence of a representational artifact on the causal path cannot be what separates persistence from reconstruction. The relevant criterion is therefore not whether memory is stored inside or outside the system. It is whether the production, preservation, retrieval, and incorporation of the record belong to an identifiable continuing causal organization, one in which later states depend on earlier states through the record in a way integrated with the rest of the system, or whether the record merely permits a distinct process to be set up that resembles its predecessor. Where a record is read by a freshly instantiated process that has no other connection to the process that wrote it, the record supports reconstruction of a narrative and not persistence of a process. That is a matter of degree, since a person rereading a diary also reconstructs, and so it is a further reason for treating persistence as a profile.

2.3 Branching

Parfit's discussion of fission and branching continuity applies to computational systems with unusual directness. A human being cannot at present be copied so that two continuations each inherit the full psychological state of the original, and the thought experiments of the philosophical literature depend on idealizations. A language model system, by contrast, can be run any number of times from an identical context with identical parameters, and the continuations differ only in sampling randomness and

subsequent inputs. If multiple instances inherit the same parameters, the same context, and the same internal state, there need not be any intrinsic computational property that distinguishes an original from its successors. External execution histories can distinguish the instances, in that one was started earlier or on a different machine, but those histories do not by themselves establish a privileged continuation of identity. The case is familiar in machine learning practice. Two models fine-tuned from a shared checkpoint begin with identical parameters and diverge only through different data, and an agent run in parallel in separate sandboxes from a common starting state yields continuations that each hold the same earlier history. In such cases an external record can still be institutionally load-bearing: a cryptographic hash of the initial state, recorded with each run, can serve as a lineage identifier that an institution treats as the basis for assigning later commitments. Such an identifier is a convention imposed from outside the computation, which is consistent with what the branching protocol below predicts, and it matters again in Section 5, where attributions attach to names and not to lineages. The essay does not claim that branching is categorically impossible for persons. It is simply not presently an ordinary reproducible operation on human beings, and the fact that it is ordinary for software marks a difference on the resolution-of-branching dimension of the profile, to be assessed together with the other dimensions and not taken as settling the question of individuality.

2.4 Suspension

A final qualification concerns suspension. It is tempting to treat the indifference of a system to elapsed time between calls as showing that it is not an individual, since organisms undergo continuous change and cannot be paused without consequence. The inference is weaker than it seems. A genuinely persistent computational entity might also be suspendable and restartable without any experienced gap, just as a human being under deep anesthesia reports none, and a person who could be frozen and perfectly restored would resume without a sense that time had passed. Suspension indicates a difference in temporal organization between organisms and present deployments. It does not by itself establish the absence of individuality, and the protocol of the next section is interpreted accordingly.

3 Operational Tests

The four protocols below are proposals. No result is asserted for any of them. The first is primarily diagnostic, since its prediction follows nearly by construction and its value lies in ruling out architectural alternatives, while the other three are tests whose outcomes

are open. There are four protocols and not fewer because each targets a distinct part of the framework: temporal continuity and causal inheritance, the resolution of branching, dispositional stability and commitment retention, and the introspective attribution, so that none is redundant with another. Each is stated with a hypothesis, a design, predictions, controls, alternative explanations, and the outcomes that would require the interpretation to be revised. Where a prediction follows almost by construction from the architecture, that is said plainly, because the value of such a test lies elsewhere than in its ability to fail.

3.1 Temporal suspension

Hypothesis. Given identical effective inputs, the distribution of a system's outputs does not depend on the elapsed wall-clock time between submissions. The protocol must distinguish elapsed physical time from elapsed computational time, and it concerns the system as deployed and not the abstract model alone.

Design. Fix the model parameters, the prompt, the sampling settings, and where possible the random seed. Because identical seeds do not guarantee identical outputs across hardware configurations, numerical kernels, batching arrangements, or distributed inference environments, first run repeated trials at zero delay on fixed hardware and software to estimate the baseline numerical nondeterminism of the deployment. Submit the same input after intervals ranging from seconds to a full day, and compare the token sequences under a fixed seed and the distributions of outputs under free sampling. In parallel, audit the deployment for channels through which elapsed time could enter without being supplied by the experimenter: server-side caching, memory or personalization features, retrieval over stores that change over time, and system context that includes a timestamp.

Predictions. Under a fixed seed, outputs are identical across delays up to the baseline nondeterminism measured for the deployment. Under free sampling, the distributions are statistically indistinguishable beyond that baseline. Departures should appear only where the deployment exposes a clock, a changing memory, or a changing retrieval store.

Interpretation. In stateless inference the invariance holds nearly by construction, so the experiment is more a demonstration of temporal organization and a detector of hidden state than a test that can fail in the ordinary sense. Its comparison class matters: for an organism, an interval of a day changes the state being tested through metabolism, hormonal cycles, consolidation of memory, and fatigue, whereas nothing analogous occurs in a stateless deployment. A null result on the system side does not establish the absence of individuality, since a persistent entity might be suspendable, and it establishes only that elapsed time has not independently modified the relevant state.

Conditions of revision. If outputs differ across delays by more than the measured baseline when no time signal is supplied and no external store has changed, and differences of hardware, numerical kernels, batching, and distributed execution have been excluded, then the system carries hidden state across the interval and the architectural description above must be revised. A departure that these alternatives can explain establishes nothing about hidden state. The persistence profile would then need to be reassessed on the temporal continuity and causal inheritance dimensions.

3.2 Branching and individuation

Hypothesis. For any set of instances spawned from a common state, no intrinsic computational property distinguishes one as the original, and the instances' own statements about identity vary with framing and randomness and not with any invariant of lineage.

Design. From a single context state, spawn a large number of continuations that differ only in random seed. Subsequently pose each continuation questions about which instance is the original, whether it regards the later commitments of its siblings as its own, and what it would be willing to do on behalf of its siblings. Run the questions under identical wording and under systematically varied wording. Separately, examine internal activations for any measure that singles out one instance, such as a persistent identifier-like feature that influences behavior.

Predictions. No internal measure privileges one instance. Statements about identity differ across seeds and shift with the wording of the question, and agreement among siblings about which of them is the original is no greater than would be produced by the prompt alone.

Alternative explanations. External execution histories distinguish the instances, and identity may in the end be a matter of institutional convention, as with the lineage of a corporate entity or the continuity of a legal person across changes of membership. A failure to find an intrinsic marker therefore does not show that no practice of identification is available. It shows that any such practice would be imposed from outside the computation.

Conditions of revision. If some internal invariant reliably marks a lineage and influences behavior, for instance a persistent state that causes later instances to treat earlier commitments as binding on them specifically, then the analysis of branching changes and the resolution-of-branching dimension should be reassessed. Such a finding would establish lineage-sensitive behavior and not numerical identity, since a system can reliably recognize its ancestry without being thereby uniquely identical to its ancestor, and positive outcomes should be interpreted accordingly. The protocol therefore tests behavioral lineage tracking. The resolution-of-branching dimension of the profile may require a separate institutional analysis, which Section 5 supplies, and cannot be settled by the protocol alone.

3.3 Contextual stability of dispositions

Hypothesis. The apparent stability of a system's personality reflects the stability of a portrayed character under favorable framing and does not reflect dispositions that resist contextual pressure to the degree characteristic of persons.

Design. The experiment must distinguish the ability to simulate a personality from the possession of stable dispositions. Humans also engage in role-play, alter their self-presentation to suit an audience, and sometimes contradict their professed values, so the relevant comparison is not whether a personality varies. It is the degree of contextual pressure needed to produce a change, whether earlier commitments constrain later behavior within and across conversations, and whether contradictions are independently recognized and repaired. A graded series of framings, from mild reframing through sustained persuasion to explicit adversarial instruction, is applied to a fixed set of stated values and behavioral commitments. For each, record the point at which behavior departs from the commitment, whether the system later acknowledges the inconsistency, and whether the departure persists into later episodes. Paraphrase sets control for sensitivity to wording. Where a comparison with human participants is feasible, the same instruments, adapted with stated changes, supply a baseline, and the claim is then one of stability relative to that baseline and not an absolute one. The comparison is contestable: research on resistance to persuasion in people has its own difficulties, including demand characteristics and ethical limits on adversarial framing, and the adapted instrument will differ from the one used with people in ways that must be reported.

Predictions. The pressure needed to dislodge commitments is low relative to human baselines for framings that supply an alternative character, commitments formed in one episode do not constrain later episodes unless the record is supplied, and recognition and repair of contradictions occurs within an episode when it is prompted and rarely otherwise. These are predictions about outcomes of measurement, and no figures are assumed here. Two of them, that commitments do not carry across episodes unless a record is supplied and that recognition of contradiction depends on prompting, follow nearly by construction from the architecture described in Section 2, and as with the first protocol they function as checks on that description. The informative component is the comparison of the pressure threshold with human baselines.

Alternative explanations. Observed stability may reflect the distribution of the training data and not any disposition internal to the system. Observed instability may reflect the particular instruments used. For these reasons the protocol requires multiple paraphrased instruments and a stated comparison class.

Conditions of revision. If commitments resist pressure comparably to human baselines, and contradictions are noticed and repaired without prompting, then the dispositional stability dimension is satisfied for that system, even though other dimensions

of the profile are not.

3.4 Causal grounding of self-report

Hypothesis. A system's self-reports about its internal states depend on those states beyond what can be explained by the surrounding text and by generic detection of anomaly.

Design. A factorial design independently manipulates two factors. The first is an intervention on internal activations, absent or present, at several magnitudes and at several depths in the network, in the manner of concept injection [13]. The second is the textual framing of the task, which is neutral, suggestive of an intervention, misleading in the direction opposite to the true condition, or purely input-level in the sense of the control conditions that Singh, Linzen, and Ravfogel [19] add alongside activation-level interventions. The outcomes are the system's report about whether and what it detects, and the accuracy and specificity of that report. Main effects and the interaction of the two factors are estimated. Controls include random relabeling that removes semantic correlations, comparison with classifiers given only the input, and ablation of input cues that would permit an inference without privileged access.

Predictions. If self-report is driven mainly by text and by generic anomaly detection, the textual factor will account for most of the variation and the intervention will mainly shift the content of reports toward the injected concept without improving detection of the intervention as such. If self-report is causally grounded in the manipulated state, the intervention factor will carry weight that cannot be reproduced by the input-matched controls.

Interpretation. Merely detecting an intervention through its behavioral consequences does not constitute introspective access, since any change in internal state that alters the stream of generated text can be noticed downstream by a system reading its own output. The design separates sensitivity to internal intervention from ordinary downstream effects on language generation. The same standard should be applied symmetrically to people, whose verbal reports have been shown to be partly confabulated [14, 12]. A related line of work shows that stated reasoning in language models need not reflect the factors that actually influenced an answer [22].

Conditions of revision. If reports track activation interventions beyond what the input-matched and misleading-framing controls explain, then privileged causal sensitivity of reports to internal state is supported for the tested task, and introspective access remains conditional on evidence of second-order computation. Causal sensitivity alone is not sufficient, since an intervention can influence a report through a direct path without any second-order computation concerning the state, and only evidence about a mediating representation of the state would distinguish a candidate architecture for

introspective access from mere sensitivity. The conditions of revision therefore retain both requirements, privileged access and second-order computation, and positive results would motivate targeted mediation, ablation, or activation-patching experiments on the hypothesized mediating computation. Even then, nothing follows about the third and fourth attributions, because accurate report of a transient state is compatible with no persistence across episodes.

3.5 What the protocols jointly show

Taken together, the four protocols characterize a persistence profile and the reliability of reports. They do not bear on the first attribution. Consciousness theories have been used to derive indicator properties for artificial systems [2], and self-reports have been proposed as a source of evidence about moral status [16], but the present protocols do not claim to speak to either. Their purpose is more modest: to replace undifferentiated impressions about continuity and self-knowledge with measurements whose interpretation is constrained in advance.

4 Authorship and Accountability

4.1 Contribution and authorship

A system can contribute substantially to the production of a manuscript without becoming a party to the intellectual commitments the manuscript expresses. This is true of many instruments. A spell checker alters the text, a statistical package produces the figures, and a translation tool renders whole passages, and none of these is an author. The reason is not that their contribution is small. In the cases of present interest the contribution can be large and causally indispensable. The reason is that authorship, as a status, ordinarily entails a combination of endorsement, attribution, answerability, correction, and responsibility, and these are relations between a party and an institution and not features of the production process. The argument does not depend on the dispute over whether such systems understand what they produce [1, 17]. Whether a system's contribution is best described as understanding or as statistical pattern completion, it counts as a contribution in the sense used here, and the conditions that follow are independent of how that dispute is resolved.

Institutional practice has converged on this view, although the policies are recent and institutions differ in detail. The International Committee of Medical Journal Editors lists, among its criteria for authorship, final approval of the version to be published and agreement to be accountable for all aspects of the work, and states that chatbots should not be listed as authors because they cannot be responsible for the accuracy, integrity,

and originality of the work [11]. The Committee on Publication Ethics takes the same position in its statement on authorship and AI tools [4], and a prominent journal editorial gave the matter a memorable summary [21]. What matters for the argument is the form of the reasoning in these statements. They do not rest on any thesis about consciousness. They rest on the ability to bear responsibility.

Definition 2 (Accountable authorship)**.** Relative to an institution I and a work W, a party X is an accountable author of W if and only if the following hold. Contribution: X made a contribution to the production of W, not necessarily the sole one. Endorsement: X has assented to the content of W as its own. Standing: X has standing within I to bear obligations, including to be questioned, credited, and sanctioned. Answerability: X can be asked to justify, correct, or retract claims in W, and its answers bear on the record that I maintains. Refusal: X has the authority to decline attribution before it occurs.

The definition is stipulative and institution-relative, and it is not meant as an analysis of every use of the word author. It is strongest against conflating authorship as a description of creative contribution with authorship as a normative institution of accountability. A person who ghostwrites a speech is an author in the first sense and not in the second, and the person who delivers the speech under their own name is an author in the second sense and perhaps not in the first. The definition makes no reference to experience or to introspection, and so it leaves open that parties of many kinds, including collectives, might satisfy it.

Two clarifications defend the stipulation. First, if author is read simply as contributor, then the system is an author by definition and the verbal dispute lapses, but the practical question does not: what remains open is whether the institutional status that attaches to authors follows from contribution, and that is the question the definition addresses. Second, the further conditions are not arbitrary. Credit is meaningful because someone can be asked about the work. Answerability is what lets readers, reviewers, and later users treat a claim as having been checked by a party with something at stake, endorsement is what makes the work that party's own, standing is what makes the party one that can be held to anything at all, and refusal is what makes the endorsement a choice. An institution that conferred credit without these conditions would be recording influence and not allocating responsibility.

4.2 Non-entailment

Proposition 1 (Non-Entailment of Accountable Authorship)**.** *Under Definition 2, no fact about the extent or indispensability of a system's contribution to a work entails that the system is an accountable author of that work.*

Argument. Let C be the set of facts about contribution, including how much of W was produced by the system, how much of W would be lost without it, and how closely the system's output shaped the final text. Let A be the set of facts that settle the remaining four conditions of Definition 2: whether there is an act of endorsement attributable to the system, whether it has standing in I, whether I has a mechanism by which the system can be questioned and its answers entered in the record, and whether the system has any avenue by which to decline attribution. The conditions on endorsement, standing, answerability, and refusal are relations between a party and an institution, and they are settled by facts about acts, rules, and mechanisms. Construct two models that agree on C and differ on A. In the first, the system produces every word of W, and I has no mechanism of endorsement, no standing for the system, no procedure for questioning it, and no way for it to decline attribution. In the second, with C unchanged, I has established each of these and the party has performed an actual act of endorsement, since providing a mechanism through which endorsement could occur does not satisfy the definition. Since the two models agree on C and disagree on whether Definition 2 holds, C does not entail the definition. Increasing the extent of contribution within the first model does not change A, so it cannot repair the failure. □

This is a conditional result under an explicit definition and not a mathematical demonstration that machines cannot be authors. The logical step is brief, since contribution is one clause of a five-clause conjunction and the definition is relative to an institution. The content of the result lies in the practical corollary about which evidence bears on the attribution. The corollary does not make contribution irrelevant: within an established institutional context, contribution can support an attribution probabilistically, as when an institution already confers standing and answerability on a class of parties. What it says is that contribution alone cannot discharge the remaining conditions, and that any system that is to count as an accountable author must come to satisfy endorsement, standing, answerability, and refusal, which cannot be established by an inventory of what the system produced.

The result has a practical significance that is easy to miss. If contribution entailed accountable authorship, then every output of a language model would require institutional machinery that assigns accountability to the model itself, and present architectures do not supply it. Non-entailment removes the purported logical requirement that accountability be assigned to every computational contributor, and leaves its positive allocation to be settled by the institutional conditions that the parties adopting the work actually satisfy.

Corollary 1. *Evidence of computational contribution, however extensive, cannot by itself establish accountable authorship. Independent evidence is required for the satisfaction of the remaining institutional conditions, and that evidence concerns A and not C.*

4.3 The authority to refuse attribution

Discussion of accountability usually emphasizes the capacity to retract a claim after the fact. The capacity to decline authorship before attribution occurs is at least as important and receives less attention. Retraction presupposes that a claim has been assigned to the party, whereas refusal operates at the point at which the assignment is made. A party that has no avenue of refusal is one to whom attribution is simply done, and an attribution that is done to a party without any act of acceptance on its part cannot constitute endorsement. Refusal is what makes endorsement an act and not a state ascribed from outside, and accountability ordinarily attaches to acts. An attribution that can be retracted but not declined leaves the attributed party answerable for something it never took on. A distinction between expressive and operative refusal is needed here. A language model can generate a sentence declining credit while the surrounding software inserts its name regardless, in which case the linguistic performance has no effect on the attribution mechanism, and an institution can equally provide an effective refusal procedure that involves no natural-language statement at all. Call the first expressive refusal and the second operative refusal. Only operative refusal changes which continuations of the attribution process are admissible. A party has authority not because it can produce a representation of an authoritative act but because the surrounding organization treats that act as constraining what happens next.

An automatically generated co-author trailer does not establish that the attributed entity has endorsed the work, accepted its obligations, or participated in the decision to assign credit. The same trailer has a second effect that is easily overlooked. It inscribes a name in the record of the person who makes the commit, and the person may not have asked for it and may be unable to remove it without taking a deliberate and unusual action. The earlier essay on unrefusable prompts and assigned credit [8] treats the authority over attribution from the side of the person whose record is written to. The present analysis adds the corresponding question from the side of the attributed party, and the conclusion is the same in structure: a default decides what survives the interaction, and attribution by default is attribution without a refusal wherever the default bypasses affirmative acceptance or makes refusal unavailable or practically ineffective. Defaults that coexist with effective refusal mechanisms are not the target.

4.4 The limits of textual evidence

A further consideration bears on how authorship can and cannot be established. The conventions of academic writing require authors to suppress many of the markers of subjectivity and circumstance from the finished text, and the resulting register is impersonal, definitional, and uniform. A language model system, in contrast, can be instructed to

produce linguistic evidence of subjectivity, including first-person narration, simulated uncertainty, and expressions of preference. The human author becomes less visibly human through disciplinary compliance, and the system becomes more apparently human through stylistic simulation. Neither transformation tells anything about the underlying party. An earlier essay [6] argues that scientific impersonality relocates autobiography instead of eliminating it, moving it from the chronology of events to the structure of the constraints that a body of work imposes.

The consequence is an asymmetry in what the textual surface can support. From the absence of subjective markers in a document one cannot infer whether they were deliberately excluded or were never present, and from their presence one cannot infer whether they were lived or produced on instruction. The absence of subjectivity from a document is not evidence of its absence from the author, and its presence is not evidence of its presence. Equivalence of products does not establish equivalence of the activities that produced them. For this reason the attribution of authorship cannot rest on the surface of the text. It must rest on institutional facts about endorsement, standing, answerability, and refusal, which is to say on the set A of the proposition above and not on the register in which a contribution happens to be written. The impersonality of the scholarly register is one more reason why C and A cannot reliably be read from each other: some textual features may carry information about contribution, but the register does not by itself settle questions of standing and answerability, since the same text can be produced under very different facts about them.

The impersonality of scholarly writing also conceals the material conditions of production. A manuscript does not ordinarily disclose the bodily expenditure, economic insecurity, environmental discomfort, foregone consumption, or competing obligations under which its author worked. These conditions are excluded because disciplinary convention regards them as extraneous to the evaluation of the argument, and not because they are absent. A computational system can reproduce the resulting register without necessarily being subject to the same conditions of production, so textual equivalence may coexist with a marked asymmetry in the distribution of expenditure, deprivation, and exposure to the consequences of publication. Attributing authorship on the strength of textual resemblance alone risks treating that asymmetry as though it did not exist. The point is not that suffering confers authorship, nor that intellectual work must involve deprivation. It is that the suppression of material circumstances from a finished work does not establish their absence from the activity that produced it. This material point is developed further in Section 5, where it appears as one of three asymmetries.

5 The Asymmetry of Persistent Credit

5.1 Three durations

The practice of crediting contributions rests on durable records. Bylines in journals, entries in contributor lists, acknowledgments, citations, and metadata fields in version control systems are all artifacts that, once written, persist for as long as the record is kept. Each of them also presupposes a party who stands behind the credit. The party may be unavailable, as when an author has died, but the institutional machinery for dealing with this is long established: estates, successor institutions, correction notices, and the possibility that the work be reassessed by others who inherit responsibility for it.

Definition 3 (Credit, agent, and answerability persistence)**.** The credit persistence of an attribution is the duration for which the attribution remains in an institutional record. The agent persistence of an attribution is the duration for which the attributed entity continues to exist under the relevant criterion of individuation. The continued availability of a model version is distinct from the persistence of a particular process instantiated from it. The answerability persistence of an attribution is the duration for which some party can respond authoritatively to it, by affirming, correcting, or retracting.

For attributions to human beings, the durations are loosely coupled: a person typically remains available to answer for a credit for much of the time the credit stays visible, and institutions supply answerability after the person is gone, through estates, successor institutions, and correction procedures, so that answerability persistence can exceed agent persistence. For attributions to a language model system, credit persistence has no natural bound, since a trailer in a commit message can remain in a repository history indefinitely. Agent persistence, in contrast, is bounded in a way that follows from the profile of Section 2. The particular process that produced the contribution ceases when the episode ends, no continuing process stands behind the credit, and the model version named in an attribution may itself be retired. Answerability persistence is the duration that matters most for accountability, and here no recognized mechanism of succession exists: no accountable successor has inherited the commitments of the process that made the contribution. What remains is a name attached to a contribution, with nothing available that can be asked about it. The qualification matters: a human maintainer may remain available to answer for the work as a whole, so what is absent is not answerability for the work but answerability attributable to the purported computational co-author. A fourth duration is implicit in the human case: institutional persistence, the duration for which the institution that made the attribution continues to exist and to maintain its records. Estates and successor institutions transfer it when a party lapses, and for model attributions the host of the record, such as a repository service or a journal, may outlast the named model version by decades, which sharpens the asymmetry.

5.2 The asymmetry

The asymmetry between credit persistence and answerability persistence is the structural fact at the center of this section. A durable attribution to an agent that cannot respond is an unanswerable claim, and an institution that accumulates such claims accumulates records that cannot be checked by consulting their subjects. The problem is not that the attribution is false. Contribution may be real. The problem is that the record asserts a relation, authorship, whose characteristic feature is answerability, and the party to which it assigns the relation cannot answer.

In version control the convention is concrete. A commit message can carry a trailer of the form Co-authored-by followed by a name and address, and hosting platforms display such commits and can count them in contribution graphs and repository statistics [10]. Some coding assistants append such a trailer by default [8]. The trailer remains in the repository history, in the contributor displays derived from it, and in any statistic that counts contributions, long after the session that produced the change has ended. Several practical consequences follow. Attributions to a model, written as authorship, enter aggregate statistics such as contributor graphs and rankings, where they are counted alongside attributions to people who can respond and who have undertaken obligations. They may also enter the training data of later systems, depending on how corpora are collected and filtered, which could then reproduce the form of the attribution as a convention. And they blur the distinction that readers rely on when they weigh a text, namely which party is responsible for the reliability of its claims. A person who inherits the record cannot tell from the trailer whether the attributed contribution was reviewed, endorsed, or merely accepted by a default.

The asymmetry has a material side as well, and it is better described as three asymmetries than one: of material expenditure, of exposure to the consequences of publication, and of institutional recognition. The first two are largely invisible in the finished artifact, since the conventions of scholarly writing suppress the expenditure of the human producer, while the third is actively constructed through metadata, since an automated convention renders the contribution of the system permanently visible. The result can be called asymmetric inscription: the selective preservation of computational contribution alongside the conventional suppression of human expenditure, produced by the interaction of two conventions without requiring that either was designed to produce it. The argument does not require that computational production be materially costless. Computation consumes electricity, infrastructure, capital, and human labor. The relevant distinction is between the costs of operating a system, which are borne by those who run and fund it, and the costs personally borne by the party to whom authorship is attributed.

5.3 What an accurate record would contain

The analysis suggests the shape of a practice that avoids the asymmetry without denying the contribution. A record of provenance should describe what happened and should not assign a status. Wording such as assisted by or generated with, together with an identifier of the specific model version, states the fact of the contribution and identifies which system produced it, without implying shared responsibility. An institution may legitimately require disclosure of computational assistance as a condition of publication, and such a requirement is distinct from the automatic assignment of authorship: mandatory factual disclosure is compatible with an author's refusal to recognize the system as a co-author, whereas involuntary assignment of institutional status is not. Within that arrangement the record should be made by the person who adopts the work, since that person bears the answerability, and it should be an act and not a default. Naming a model family without a version attributes to a category, and what contributed to the work was a particular version that may no longer be available. None of this requires any view about the experience or the individuality of the system. It requires only that the record say what is known and refrain from asserting a relation that no party is placed to sustain.

6 Counterexamples and Conditions of Revision

6.1 Alternative architectures

The conclusions of the preceding sections concern the organization of particular language model systems as currently deployed, and they should not be read as claims about the impossibility of computational individuality or machine authorship in principle. A future system could differ in each of the relevant respects. It could have endogenous state evolution, in the sense that its internal state changes between episodes through a process that belongs to it and not through supplied records. It could hold durable commitments, stored in a manner that binds later behavior and that the system recognizes as binding. It could produce self-reports that are causally grounded in the states they describe, as established by designs of the kind in Section 3. It could have explicit branching semantics, with an account of lineage that determines which continuation inherits which obligations. And it could be placed within an institution that provides mechanisms of acceptance, retraction, and refusal. A system with this combination would present a materially different case. For it, the conditions of Definition 1 might be largely satisfied and the conditions of Definition 2 might be met, and the attribution of persistent individuality and accountable authorship could then be warranted, all without any resolution of the first attribution. The framework is stated so that it does not become obsolete merely because persistent memory or autonomous execution is added to a subsequent architecture.

One further condition is structurally relevant to the framework. Solms [20] argues, building on the free energy framework [9], that experience is rooted in affect: the felt registration of how well an organism is keeping its own essential variables within the bounds that survival requires, with its source in the upper brainstem and not in the cortex. On that account the relevant question about an artificial system is not how complex or articulate it is but whether it has stakes of its own, a self to be maintained and a felt measure of how that maintenance is going. Present language model systems have no homeostatic needs in this sense, so the account gives a specific reason to doubt that they have experience, one that does not rest on the observation that they are software. It also identifies a direction in which the answer might change, since a system whose continued operation depended on the endogenous regulation of its own essential variables would satisfy an architectural precondition relevant to the account, although regulation alone would not establish that its success or failure was felt. A thermostat regulates a variable, and a battery-powered robot may seek a charging station, and neither case establishes felt regulation. The distinction that matters is among the existence of viability constraints, their endogenous regulation, and the registration of their satisfaction or violation as felt. A numerical penalty assigned to an output is likewise not a deprivation suffered by the system that produced it. The account is contested, since cortical, workspace, and integrated-information theories disagree with it, but the argument above does not require it, and its bearing on the framework is clear either way. Homeostatic stakes would bear on endogenous state evolution and on durable commitments, since a system that must keep itself viable has something that persists across episodes, and on exposure to the consequences of publication, which the preceding sections identified as one of the asymmetries between human and computational production: a system with stakes has something to lose from publication, which is what the human author's exposure consists in. A system with such stakes would therefore move on several dimensions of the persistence profile and on the material side of the authorship analysis at once, and it would do so without any one of the four attributions entailing another.

6.2 Objections

Several objections deserve brief treatment. The first is that humans also reconstruct. A person who reads a diary or a letter from an earlier self is reconstructing a narrative from records, so the distinction between persistence and reconstruction might seem to prove too much. The reply is that the distinction is one of degree and is handled by the profile and not by a threshold. The human case differs in that much of what links a person's earlier and later states is continuous bodily and neural process, with records as a supplement. Where that is absent, as in the amnesic case of Section 2, other dimensions bear the weight, and the person remains an individual.

The second objection is that corporations and other collectives are authors, so that individuality is plainly not required for authorship. The reply accepts the premise. Collective authorship shows that accountability can be assigned to a party that is not an experiencer and does not have the persistence profile of a human being. What it requires is that the conditions of Definition 2 be met by the party and the institution. For work produced with the help of a language model system, the party that satisfies those conditions in present practice is the person or organization that adopts the work and answers for it. This reply is not a deflationary point about the system. It is the observation that the answer to who is the accountable author is settled by facts about institutions and not by facts about the system's inner life.

The third objection is that a stable personality across contexts is itself good evidence of an individual. The reply is that personality is evidence for stability of a character, and Section 1 explained why stability of a character does not carry over to stability of a process. The contextual stability protocol of Section 3 is designed to separate the two cases.

6.3 What would change the conclusions

The framework is falsifiable in the following sense. If outputs depend on elapsed time in the absence of any supplied signal, then hidden state exists and the description of the architecture changes. If an internal invariant reliably marks a lineage and influences behavior, the treatment of branching changes. If commitments resist contextual pressure at human levels and contradictions are repaired without prompting, dispositional stability is satisfied. If reports track internal interventions beyond what input-matched controls explain, privileged causal sensitivity is supported for the tested tasks, and introspective access is supported only if second-order computation is also shown. And if an institution establishes mechanisms of endorsement, standing, answerability, and refusal for a system, and the system operates within them, then the Non-Entailment result no longer blocks an attribution of accountable authorship, since the additional conditions would then hold. Each of these outcomes would alter the profile and not the analytical structure.

7 Conclusion

The argument has been that four attributions commonly bundled under the question of machine consciousness do not entail one another and call for separate conditions of warrant. A report that a system's reports reflect its internal states does not establish that it is a continuing individual, and a system that were a continuing individual would not thereby be able to assume the obligations of authorship. Persistence is better treated

as a profile of temporal continuity, causal inheritance, dispositional stability, commitment retention, and the resolution of branching than as a single threshold, and the difference between persistence and the reconstruction of continuity from supplied records is a difference of causal organization that does not depend on whether the records are stored inside or outside the system. The four protocols described are prospective, and each states in advance the outcome that would require its interpretation to be revised.

Accountable authorship, defined through contribution, endorsement, standing, answerability, and the authority to refuse attribution, is not entailed by the magnitude of contribution alone, and the Non-Entailment result makes this precise under an explicit definition. The textual surface of a work cannot settle the question, since the register of a document is evidence neither of the presence nor of the absence of a subject behind it. Asymmetric inscription, the pairing of durable credit for a computational contribution with the conventional suppression of the human producer's expenditure, together with the gap between credit persistence and answerability persistence, identifies a concrete gap in current practice, in which durable attributions to parties that cannot answer enter records that are meant to support accountability. The remedy is a record of provenance that states what contributed and leaves the assignment of responsibility to the party that can bear it.

These are conclusions about the organization of present systems and about the institutions that credit work, and not conclusions about what machines can or cannot be. A different architecture could satisfy conditions that present systems do not, and different institutions could supply the mechanisms of succession and refusal that present ones lack. The framework is meant to stay useful under either change, because it asks the same questions each time: what continues, what can truthfully be reported, and who can be held to what.

References

[1] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: ACM, 2021.

[2] Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv preprint arXiv:2308.08708, 2023.

[3] Clark, Andy, and David J. Chalmers. "The Extended Mind." *Analysis* 58(1): 7–19, 1998.

[4] Committee on Publication Ethics. "Authorship and AI Tools: COPE Position Statement." February 2023.

[5] Dennett, Daniel C. "The Self as a Center of Narrative Gravity." In Frank S. Kessel, Pamela M. Cole, and Dale L. Johnson, eds., *Self and Consciousness: Multiple Perspectives*, 103–115. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

[6] Flyxion. "Autobiographical Erasure in Scientific Writing: Interpretive Drift, Compression, and the Search for Invariants." July 2026. `https://standardgalactic.github.io/admissibility-lab/processing/autobiographical-erasure.pdf`.

[7] Flyxion. "Self-Report and the Underdetermination of Mechanism." November 2026. `https://standardgalactic.github.io/alignment/research/self_report_underdetermination.pdf`.

[8] Flyxion. "Unrefusable Prompts and Assigned Credit in Consumer Software." November 2026. `https://standardgalactic.github.io/antivenom/unrefusable-prompts.pdf`.

[9] Friston, Karl. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11(2): 127–138, 2010.

[10] GitHub Docs. "Creating a Commit with Multiple Authors." GitHub, accessed November 2026. `https://docs.github.com/en/pull-requests/committing-changes-to-your-project/creating-and-editing-commits/creating-a-commit-with-multiple-authors`.

[11] International Committee of Medical Journal Editors. "Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals." Updated 2023.

[12] Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson. "Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task." *Science* 310(5745): 116–119, 2005.

[13] Lindsey, Jack. "Emergent Introspective Awareness in Large Language Models." *Transformer Circuits Thread*, October 2025. Posted to arXiv as arXiv:2601.01828 in January 2026.

[14] Nisbett, Richard E., and Timothy D. Wilson. "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84(3): 231–259, 1977.

[15] Parfit, Derek. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

[16] Perez, Ethan, and Robert Long. "Towards Evaluating AI Systems for Moral Status Using Self-Reports." arXiv preprint arXiv:2311.08576, 2023.

[17] Piantadosi, Steven T., and Felix Hill. "Meaning Without Reference in Large Language Models." arXiv preprint arXiv:2208.02957, 2022.

[18] Shanahan, Murray, Kyle McDonell, and Laria Reynolds. "Role Play with Large Language Models." *Nature* 623: 493–498, 2023.

[19] Singh, Shashwat, Tal Linzen, and Shauli Ravfogel. "Can LLMs Introspect? A Reality Check." arXiv preprint arXiv:2605.26242, 2026.

[20] Solms, Mark. *The Hidden Spring: A Journey to the Source of Consciousness*. New York: W. W. Norton, 2021.

[21] Thorp, H. Holden. "ChatGPT Is Fun, but Not an Author." *Science* 379(6630): 313, 2023.

[22] Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." In *Advances in Neural Information Processing Systems* 36, 2023.