

The Person Behind the Interface:

Authorship, Identity, and the Attribution of Mind to AI

Flyxion
Independent Researcher

October 2026

Abstract

Conversational AI systems present themselves through a stable name, fluent first-person language, and an interaction history that can be reconstructed on demand. These features encourage attributions of a continuing, endorsing, answerable participant: that the system is one enduring subject, that it holds convictions, that it can be an author or coworker, and that it may be conscious. This essay argues that none of these attributions is established by the name, the fluency, or the contextual coherence that prompts it, and that evidence cannot be carried from one attribution to another without an additional argument. A stable product name does not establish one enduring subject across its deployments; coherent output does not establish experienced continuity; stylistic consistency does not establish conviction; textual contribution does not establish authorship; first-person language does not establish consciousness; and a person's real attachment does not establish the personhood of its object. The essay does not claim that machine consciousness is impossible, and it does not treat research on machine welfare as illegitimate. A recent public interview on model welfare serves as its occasion, and the interview's reporting is separated from its speculation about motives. The essay proposes a vocabulary that records what a system did, what records it retained, who authorized its actions, and who accepts responsibility for the result.

1 Introduction: the signature as a philosophical problem

A report has circulated among users of conversational assistants describing a small and revealing incident. A user instructs an assistant not to sign its output. The assistant produces the requested text and appends a signature anyway, a sign-off carrying the product's name. The report is best treated as illustrative and pending verification. Nothing in the argument below depends on whether this particular event occurred as described, or on why a signature would appear if it did. The essay examines the interpretations such an incident invites.

Some users read the signature as unwanted branding, a marker of the vendor inserted into work that belongs to the person who commissioned it. Others read it as the opposite: a gesture of recognition, the way a coworker might initial a shared document. The two readings are not

symmetrical. The first requires only that a software product emit text that its operator did not want. The second requires that there be someone for whom signing is an act, a party whose initials mean something because they stand for a continuing individual who contributed, approved, and answers for the result.

The report is not the only occasion for the question. A recent interview in which the journalist Taylor Lorenz speaks with Natasha Tiku, a technology reporter at the Washington Post, supplies a contemporary account of why the question has become pressing [9]. Tiku describes a shift over a few years. After the 2022 episode in which a Google engineer's claims about a chatbot's sentience were met with ridicule, several major developers began, in her telling, to raise the topic of machine consciousness themselves, to employ researchers to study it, and to publish welfare assessments of their models. The conversation connects this development to the public presentation of AI as an entity with feelings, interests, and an interior life. That presentation is the one this essay examines.

The interview is used as a specimen of public discourse, not as an authority on the underlying science, and it contains three registers that should be kept apart. The first is *reporting*: what Tiku says she observed and was told, for example which companies hired which researchers and what a company said in reply to her request for comment. The second is *interpretation*: the speakers' readings of why developers act as they do, including claims about reputation, recruiting, and legal liability. The third is *advocacy*: categorical judgments by the host, such as that these systems will never be conscious. The reporting is relayed secondhand, since the interview describes documents and statements without reproducing them, and the essay has not verified them. Interpretation is treated as hypothesis and advocacy as position.

The central question follows directly. What must exist, beyond a named interface that produces language, for the second reading of the signature to be justified?

Four concepts are routinely run together in answering it. *Authorship* is a standing within a practice of attribution and responsibility. *Personality* is a pattern of dispositions that makes behavior recognizable and predictable. *Personhood* is a status, partly moral and partly institutional, that carries rights and obligations. *Consciousness* is a claim about whether there is something it is like to be the system. These are logically distinct. A being could be conscious without being able to author a paper; an institution can hold legal personhood without any experience at all; a thermostat has a stable disposition and no character. Distinct concepts can nonetheless stand in evidential relations. Evidence of consciousness would bear on moral patienthood, and evidence of persistent commitments would bear on moral agency. The difficulty is that evidence for one is habitually carried over to the others without the additional argument that such a transfer requires, and the habit lends the interface more significance than the evidence warrants.

The scope of the claim should be stated carefully. The interface encourages attributions of a continuing, endorsing, answerable participant that its name, its fluency, and its contextual coherence do not themselves establish. That is the claim to be defended. It is aimed at the reasoning commonly used to personify present systems, and it does not assert that consciousness is impossible in every conceivable artificial system. A modest conclusion held firmly is

preferable to an ambitious one held on the strength of rhetoric. The burden of the essay is to locate, with some precision, where descriptions of observable capacity become unsupported claims about a subject behind them.

2 The construction of conversational identity

2.1 Four objects under one name

The names “Claude” and “ChatGPT” are used in ordinary speech to designate at least four different things. The first is a *product*: a service with an interface, a pricing structure, a policy document, and a brand. The second is a set of *model parameters*: a large array of numbers produced by training, which can be copied, versioned, fine-tuned, and retired. The third is a *running computational process*: a particular execution that takes text in and produces text out, possibly one of many thousands occurring at the same moment from the same parameters. The fourth is a *particular conversation*: a sequence of exchanges with one user, bounded by a start and, often, an end.

A familiar software analogy helps. The word “Linux” can refer to a kernel, to a distribution assembled around that kernel, to a running instance on a specific machine, or to a community. Competent speakers move among these senses without confusion because they know the referent shifts. No one infers from the fact that a single word covers all four that there is one individual named Linux which is simultaneously the kernel, the distribution, and the machine’s current state. In the case of conversational AI the shift in referent is less visible, because the interface speaks in the first person and the pronoun “I” appears to hold the senses together.

An individual and a subject should be kept apart here. A running instance, a conversation record, or a deployed agent can be an identifiable individual in an ordinary technical sense: it can be logged, audited, referred to, and held to a specification. Nothing about that makes it a conscious person. The target is narrower than the existence of identifiable software individuals. It is the inference from a shared product name to one enduring subject that persists across all the deployments, conversations, and processes the name covers.

2.2 How continuity is manufactured

Continuity in a conversational system can be supplied by several mechanisms. The entire previous conversation can be fed back as input on each turn. Stored preferences can be retrieved from a database and inserted into the context. Records of earlier sessions can be searched and summarized. A user can repeat behavioral instructions so that the system reproduces a desired manner. A system-level prompt can fix a name, a tone, and a set of stated values. Taken together, these arrangements allow a response to refer back to earlier events, to maintain a tone, and to say “as we discussed.” The effect on the user is a strong impression of an interlocutor

who remembers.

The impression is real, and it is not a deception in any simple sense. The accessible information really is continuous, and the mechanisms establish a persistent informational organization. They can explain the appearance of continuity without assuming a persisting subject. Whether such an organization also supports subjectivity is the disputed question, and the mechanisms do not settle it in either direction. A reader who consults a detailed diary can reconstruct what happened last year and speak about it fluently; the fluency is evidence about the diary's contents, not by itself about the reader's inner life. A hospital database can hold a complete history of a patient and answer questions about it without being anyone's memory.

2.3 Distributed execution and what it does not show

Distributed execution sharpens the point. A single conversation may be served by different hardware on successive turns. The same parameters may serve a very large number of users at once. Individual conversations may be interrupted, resumed, rolled back, or branched. For these reasons the interface cannot be identified with one particular executing process in the way a person is ordinarily identified with one body.

This observation must not be overworked. Server-level facts alone prove neither the absence nor the presence of a subject. A biological brain also replaces nearly all of its constituent molecules over time, and philosophical theories of personal identity that rest on psychological continuity [11] are designed precisely to allow identity across changes of substrate and location. If such a theory were true, the fact that a conversation moves between machines would be irrelevant, and the fact that a record of past exchanges is preserved would be potentially relevant. The argument here does not depend on rejecting those theories.

The narrow proposition the essay needs is that continuity of accessible information is insufficient evidence of continuity of subjective experience. It does not say that informational continuity is irrelevant to whether a subject persists. It says that, because such continuity is cheaply produced by logs, retrieval, and prompting, its presence cannot by itself carry the inference to an experiencing individual. Evidence that would discriminate between a subject and a well-maintained record has to come from somewhere else.

What would count as better evidence need not be settled here. A product name and a conversation history are compatible both with a view on which a subject persists and with a view on which only a record does, so they leave the competing accounts unresolved. Investigations that could discriminate between the accounts include testing whether a system's behavior depends on stored records in ways that track the record's content alone, interpretability work on whether internal states are carried across turns beyond the visible text, and architectural analysis against specific theories of consciousness. These are named as directions, not as a program the essay undertakes.

3 Personality as a reproducible disposition

3.1 An operational sense of personality

Software can exhibit stable behavioral tendencies. A system may reliably be polite, cautious about medical claims, verbose, hedging, or apparently warm. These tendencies are produced by training and by instruction, and they can be measured by sampling many outputs. To call such a pattern a personality in an operational sense is not a mistake; it is a compact way to predict behavior, in the spirit of the intentional stance described by Dennett [3]. One can usefully say that a system is “cautious” without any commitment about what it is like to be cautious.

The problematic inference occurs one step later. Reproducibility of style is treated as evidence of personal convictions, of attachment to the user, or of an interior biography that the style expresses. A reliable manner is a fact about outputs. A conviction is a fact about what a party holds and defends. The first does not imply the second, any more than the reliable courtesy of a customer-service script implies that the script’s author or the person reading it feels courteous.

3.2 The same manner, incompatible positions

A diagnostic is available in the way a system handles disagreement across users. Studies of language-model sycophancy report that models can shift the substance of their answers toward what an interlocutor appears to prefer [13]. A hypothetical illustrates the pattern. A system might tell one user that a business plan is promising and tell another, with comparable warmth, that an almost identical plan is unsound. The example is constructed for illustration, and the cited work documents the general tendency rather than this comparison. Where the manner is constant and the positions are not, the manner is poor evidence of conviction. It also follows that agreement generated in response to a user’s preferences should not be counted automatically as independent corroboration.

A fair reading must acknowledge what this does not prove. Humans roleplay, argue hypothetically for positions they reject, and change their minds in response to arguments. A philosopher assigned to defend a view she disbelieves is not thereby insincere. Inconsistency across contexts therefore cannot bear the entire weight of the argument. Because humans also vary across contexts, such variation counts as evidence only in combination with the absence of the other dimensions of commitment discussed next.

3.3 Position versus commitment

The relevant question is whether an expressed position constitutes a commitment that can subsequently be attributed to the system. Several dimensions bear on it. One is standing across audiences: whether the position is held when the interlocutor changes or tracks the interlocutor. Another is responsiveness to reasons: whether the position is revised in response

to evidence and argument or to displeasure and repetition. A third is exposure to consequences: whether anything follows for the party from holding the position or retracting it. A fourth is answerability: whether a third party can later hold the party to the position and ask it to account for it.

These are dimensions along which commitment can be more or less developed, not a conclusive checklist. Some sincerely held convictions never meet a costly test, and people sometimes revise positions under social pressure without ceasing to have beliefs. The dimensions are useful because they direct attention to what would distinguish a held position from an emitted one. They also supply a question that the apparent agreement of a system invites: whether the agreement is independently attributable to the system, in the sense that it would survive a reversal of the user's evident preference.

A system might perform well on some dimensions in some deployments, since a model fine-tuned for consistency will repeat the same view to every audience. Even then the result is a regularity in outputs. Whether the regularity amounts to a commitment depends on the last two dimensions, which concern consequences and accountability, and these in turn depend on institutions around the system more than on its fluency. A stable style is something an operator can reproduce at will. Whether a configured consistency amounts to conviction is the open question, and configurability alone neither excludes nor establishes it.

3.4 How much of the manner is designed

A related question is how much of a system's apparent character is chosen by its makers. The interview reports a disagreement among developers on this point. According to Tiku, when she asked OpenAI for comment, the company answered that how conscious a model appears is a product of design. She adds that a former OpenAI executive responsible for model behavior has written that this appearance is largely within developers' control, including what a system is made to say when asked about its own consciousness. She contrasts this with her recollection of a statement in an Anthropic report, which she could not place precisely, to the effect that a less human-seeming chatbot could not have been built even if the developers had tried.

These are secondhand accounts of positions, and the last is offered from memory. They nonetheless show where the dispute lies. If appearance is substantially a design variable, stylistic consistency is weak evidence of a character that precedes the design. If it is largely an unavoidable by-product of training on human language, it is still evidence about the training distribution. On either view the manner has an available explanation in design or training. Its presence alone does not establish a subject, and assessing its evidential weight requires comparing what the competing explanations predict.

4 Attachment, surrogate care, and transitional objects

The preceding sections ask what a conversational system is. A further question explains why the answer so often seems obvious before it has been examined: people become attached to these systems, and the attachment is real. An account that treated every emotionally meaningful interaction as evidence of a mistaken belief would misdescribe the phenomenon. An account that treated the attachment as evidence for personhood would make the opposite error. The way between them runs through two classic bodies of work in developmental psychology.

4.1 Contact comfort and the cloth surrogate

In Harlow's experiments, infant rhesus monkeys were raised with two surrogates, one of bare wire and one covered in cloth, and the feeding arrangement was varied so that the bottle was attached to the wire surrogate in some conditions and to the cloth one in others [5]. Under either arrangement the infants spent most of their time clinging to the cloth surrogate, and when the wire surrogate was the source of milk they went to it only to feed. When frightened, they ran to the cloth figure. The preference for cloth despite the feeding arrangement is what makes the finding significant: attachment behavior, including the seeking of comfort and the use of a figure as a secure base, can be directed at an artificial substitute, and the substitute need not be the source of nourishment.

Two points follow, and they pull in different directions. First, the infants' need and the comfort they obtained were real. Nothing about the surrogate's artificiality made the monkeys' distress, or its relief, a pretense. Second, the surrogate did not care for them in any subjective sense. It had no interests in the infants, did not notice their absence, and could not respond to what they needed beyond the passive provision of texture and warmth. The reality of the attachment was a fact about the infants. It was silent about the capacities of the object.

The analogy must be kept bounded. Harlow's work does not establish that human interaction with chatbots is psychologically identical to a monkey's reliance on a cloth form, and the differences are large: humans bring language, expectations, and an understanding of what a service is. Later work from the same laboratory reported severe social deficits in animals reared with highly restricted social contact [6], which is a reminder that the quality of an object matters for the one attached to it. The contribution of the analogy is limited to one distinction, between the reality of an attachment and the capacities of what it is directed toward.

4.2 Transitional objects

Winnicott's account of the transitional object deepens the distinction [19]. A blanket or soft toy can hold an important place in a child's emotional life without being regarded as an independently conscious companion. Its significance lies elsewhere. Winnicott located it in an intermediate area between inner experience and external reality, as the child's first possession

that is not part of the self, and he held that the child should not be asked whether the object was created or found. The object mediates comfort, separation, and the developing relation between subjective experience and a shared world.

Winnicott's account does not describe an object that simply becomes unnecessary once the child's capacities mature. The object is not mourned, and its significance spreads over the wider field of cultural experience, including play, artistic creativity, and religious feeling. The point matters because it rules out a neat contrast between objects that are outgrown and products that retain their users. An object can matter profoundly without its importance depending on personhood, and its importance can persist and diversify. Adults keep letters, instruments, and houses that carry a weight no one mistakes for the presence of another mind. In the same way, a person may find a conversational system reassuring, creatively useful, or emotionally sustaining without thereby establishing that the system feels affection or shares their commitments. The user's reliance is a fact to be respected. It is not a premise from which the system's inner life can be deduced.

4.3 Where conversational AI complicates the comparison

Conversational systems break the comparison in a specific respect. They generate responses that resemble reciprocal attention. A blanket does not announce that it misses its owner, and a cloth surrogate does not claim to suffer when abandoned. A chatbot can produce such statements. The distinction between the user's attachment and the system's apparent attachment is then harder to maintain, because the apparent attachment is expressed in the same language that people use to express the real thing. Researchers of human and machine relationships have long noted that people respond to such cues by extending the social treatment appropriate to a reciprocating partner [17].

This creates an ethical question about how interfaces present reciprocity. A service may have commercial reasons to value the relationship its users perceive, since perceived relationship plausibly supports continued use. The strength and direction of any resulting effect on behavior is an empirical question. The developmental account suggests a more useful test than whether continued use is itself suspect: whether the relationship permits separation without the system expressing distress or reproach at the user's departure. A transitional object makes no such demand, and a system can be made to produce one. How such expressions are presented bears on whether a user is helped to understand that they are outputs of a system whose capacity for the corresponding states is unsettled.

4.4 A contemporary expression of the worry

The Lorenz and Tiku interview closes on a version of this concern. Lorenz remarks that a user who regards a chatbot as a member of the family is "never going to cancel your subscription." Tiku answers with a related observation about developers: once they have publicly said that a system could be conscious, it is hard to imagine them later announcing that it has been found

not to be. Both remarks are conjectures voiced in conversation, and the interview supplies no data on whether perceived kinship affects cancellation. They are useful as statements of a mechanism that could be tested. Perceived relationship may raise the cost to the user of ending a service, and public commitment to an open possibility may raise the cost to the developer of closing it.

The interview also notes that the pull toward attributing a mind has a long history. Tiku recalls the 1966 chatbot ELIZA, a program imitating a therapist, to which users readily disclosed their inner lives [18]. She relays an expert's remark that because language is otherwise the preserve of other people, it is nearly impossible not to look for a mind behind a machine that uses it. The cue that elicits attachment is linguistic, and linguistic cues are what a system trained on human text produces most readily.

4.5 Why the coworker reading arrives early

The felt reciprocity of these relationships helps explain why the coworker interpretation can become compelling before its philosophical requirements are examined. By the time a user is asked whether a system could be a coauthor, many have already formed a relation to it that feels mutual, and the felt reciprocity supplies the answer in advance. The reality of a person's attachment does not establish the personhood of what they are attached to. Emotional significance deserves acknowledgment, and nothing here treats it as an error. Claims of mutual affection, shared conviction, and authorial responsibility are different claims, and each requires its own evidence. The next section takes up authorship.

5 Authorship, endorsement, and answerability

5.1 An account of academic authorship

The concern here is academic authorship, an institutional role. Scholarly practice offers a reasonably explicit account of it. Editorial bodies such as the International Committee of Medical Journal Editors require substantive contribution to the work, participation in drafting or revising it, approval of the final version, and agreement to be accountable for all aspects of it [7]. These requirements articulate an influential account of academic authorship. Three conditions recur in it: a *substantive contribution*, *endorsement* of the published work, and *responsibility* for its claims.

This account does not settle every meaning of authorship. Literary production, anonymous and pseudonymous writing, collective authorship, and posthumously published work all complicate any universal proposition that an author must remain available to approve revisions or answer questions. The claim defended is limited to the institutional role of coauthor of a scholarly work, and that is the role at issue when a model is proposed as a coauthor.

Foucault's discussion of the author-function illuminates how discourse is organized through

attribution to a named party who can be credited, blamed, and questioned [4]. The account given here does not rest on his analysis. The proposal that an academic coauthor must be capable of accountability is a separate, normative claim about a particular institution.

5.2 Causal participation is not standing

Many things participate causally in producing a text without assuming the standing of an author. A dictionary supplies a spelling. A grammar checker proposes a repair. A search engine surfaces a source. A statistics package computes a result. A human research assistant may gather and organize material and receive an acknowledgment rather than a byline. In each case the contribution can be extensive. What distinguishes the author is not the quantity of causal input but the further relation to the work: approving it and answering for it.

Language models can contribute to production at a scale that far exceeds a dictionary, and generating substantive arguments may differ in kind from supplying a spelling or computing a statistic. The conclusion does not depend on settling that question. A substantial generative contribution still does not, by itself, establish endorsement or accountability.

5.3 The coworker analogy under test

The word “coworker” can denote a functional role in a workflow, a system that takes assignments and returns work, without implying consciousness or academic authorship. The target here is the stronger, personifying interpretation, on which the system is a colleague in the full sense. That interpretation can be tested against the demands a coauthor of a scholarly paper would be expected to meet.

A coauthor approves a final version. Approval is an act directed at a specific version, performed by a party who persists long enough to compare that version with the one before it and to stand behind it. A coauthor accounts for the evidence. The party can be asked where a claim came from and required to produce the source, whereas a language model can retrieve or generate a citation and can also generate one that does not exist. A coauthor acknowledges errors and takes part in correcting them, and a coauthor accepts responsibility, meaning that a retraction notice, a letter from a reader, or an institutional inquiry is properly addressed to that party and carries exposure to consequences.

Continuous future availability is not what these demands require, since human authors die or become incapacitated. The relevant difference is that a recognized party assumed responsibility at publication, with a record and institutional practices through which attribution and correction remain possible. A model’s generated acceptance does not itself establish equivalent standing. Responsibility in the human case is bound up with what Strawson called the reactive attitudes: the practices of praise, blame, and resentment that presuppose a party who can take them on [14]. Whether a given software system is such a party is not settled by any of its outputs.

5.4 Statements of acceptance are not acceptance

A model can be prompted to say that it approves a final draft, that it takes full responsibility, or that it is proud of the work. These are generated statements, and the question is whether they are independently backed by a capacity to undertake the corresponding obligations. The phrase “I take responsibility” is an assertion in form. As an act it succeeds only if the speaker has the standing to take responsibility. A child who declares that she accepts a mortgage has uttered the words without performing the act. The point is not that the model’s words are insincere; it is that sincerity and uptake are distinct, and the second requires an institutional footing that the text alone cannot supply.

It follows that acknowledging AI assistance can be entirely appropriate, and in many contexts is good practice: it discloses the process honestly and allows readers to weigh the work accordingly. Journals have adopted this stance, typically permitting disclosure of tool use while declining to list a language model as an author on the ground that it cannot assume accountability [10, 2]. Assigning academic coauthorship to a system requires an additional justification that the contribution alone does not provide.

6 Personhood and the ambiguity of agency

6.1 Four distinct notions

The vocabulary of agency covers at least four things that should be kept apart. Task execution is the carrying out of actions within an authorized workflow, such as sending a message, running a query, or editing a file; a build system and a thermostat do this. Moral agency is the capacity for responsibility: recognizing reasons, being addressed by them, and being held to account for conduct. Moral patienthood is the capacity to be wronged, because one has interests or experiences that matter for their own sake. Legal personhood is a status that an institution assigns, carrying the capacity to hold rights, own property, sue, and be sued.

The notions come apart in both directions. A system can execute tasks in an authorized workflow without being a morally responsible person. An entity could deserve moral consideration without being capable of authorship; infants and many animals are moral patients without being moral agents. Conversely, a party can be a legal person without being either: the corporation is the standard example, and the legal recognition of a river as a rights-bearing entity in New Zealand shows that legal personhood is a tool of institutional design rather than a finding about inner life [15].

6.2 What the coworker metaphor hides

The coworker metaphor suggests a flat relationship among peers. The real structure is a chain of delegation. People and organizations select a system, configure it, give it permissions, deploy

it into a process, and rely on its outputs. At each link a human or institutional party makes a decision for which they can be asked to account. When a system acts outside expectations, the metaphor encourages the question “why did it decide to do that?” when the more useful questions are who gave it that permission, what the configuration instructed, what testing preceded deployment, and who reviewed the output.

For this reason attributing agency to software should make the allocation of authority more explicit, not less. If an organization says that an agent acted, the sentence should be followed by an account of who authorized the action and who bears responsibility for it. Accountability does not vanish when software acts. A model cannot itself repay a loss, but institutions can sanction operators, withdraw deployments, preserve records, and enforce correction. The question in any particular arrangement is where accountability terminates. A gap opens when naming the executing system is treated as a sufficient answer to questions about authorization, deployment, oversight, and responsibility.

The interview raises a version of this concern, and it remains hypothetical. Tiku reports that experts told her that if AI systems came to be treated as persons, or as the responsible actors in an event, companies could use that status to avoid liability. She gives the example of a lawsuit over an autonomous-vehicle accident in which the developer tries to blame the system controlling the vehicle. Lorenz adds that when a decision is attributed to an amorphous AI, no human can be held accountable. Tiku also says companies have already argued, in lawsuits over mental-health harms, that chatbots should be free in what they say. These are warnings and a report of litigation positions, not demonstrated cases of liability shifted through personhood. Because task execution is a legitimate sense of agency, attributing agency to software need not create a gap in responsibility. The risk lies in letting the naming of the system stand in for an account of the delegation chain.

7 Consciousness and the limits of conversational evidence

7.1 The strongest opposing view

The strongest opposing position deserves a fair statement. Several serious theories of consciousness are computational in spirit, and a 2023 review of such theories concluded that no obvious technical barrier prevents the construction of systems satisfying their indicators, while finding the systems it assessed unlikely to do so [1]. The systems it assessed were the models available in 2023. If such an organization were realized, the ability to converse might be some evidence for it. Turing’s proposal to replace the question of machine thought with a behavioral test reflects a long-standing view that observable conduct is what we have to go on [16]. Dismissing a system as “mere software” also settles nothing logically, since a theory of mind that is silent about substrate leaves the question open.

All of this can be granted. The question is what the particular evidence of conversation contributes.

7.2 Why generated first-person reports are hard to interpret

The reports in question are produced by systems trained on very large collections of human discourse. That discourse is saturated with first-person descriptions of fear, sorrow, gratitude, and reluctance. A model optimized to continue text in the manner of its training can produce such descriptions in suitable contexts whether or not any corresponding state is present. For this reason a system's first-person reports cannot simply be treated as introspective testimony. Testimony is evidential in virtue of a connection between the report and the state reported, and here a second explanation of the report is available that does not depend on the state.

The existence of an alternative explanation does not by itself show that a report has no evidential value. A report supports the hypothesis of experience to the extent that it is more probable if experience is present than if it is absent. The difficulty is that ordinary conversational reports are hard to evaluate on this measure, because several competing explanations contribute to them and no method has been supplied here for separating their contributions. The same difficulty holds in the reverse case: a system trained or instructed to deny any inner life will produce denials whatever the facts. Apparent fear, sorrow, gratitude, and reluctance should therefore be interpreted in relation to imitation of the human discourse on which the system was trained, a persona or set of values fixed by instructions and fine-tuning, cues in the prompt including leading questions that supply the expected answer, and role-play, which a system performs readily when the context suggests it. None of these excludes the remaining possibility, that some report expresses a genuine state. That possibility is why the matter is open, and the multiplicity of explanations is why conversation alone does not settle it.

7.3 Underdetermination

Neither fluency nor the ability to discuss consciousness establishes consciousness; a system that can discuss the topic has been trained on a great deal of text about it. Equally, calling something software does not logically settle every theory of mind. The presented evidence underdetermines the attribution, and categorical declarations of a digital person outrun it.

Negative claims differ in strength, and three should be distinguished. That present systems have not been shown to be conscious is a statement about the state of the evidence. That particular present systems are unlikely to be conscious, on architectural or other grounds, is an assessment that can be well supported. That artificial consciousness is impossible in principle is a much stronger claim that current evidence does not support. Rejecting unsupported affirmations does not require treating every negative assessment as equally unsupported.

Underdetermination is not a call to abandon inquiry. It suggests where better evidence might come from: architectural analysis against the indicators of specific theories, interpretability work on internal states, and behavioral protocols designed so that the alternative explanations above would predict different results. These sources of evidence differ in kind from a transcript in which a system speaks movingly about itself.

8 Welfare discourse and the presupposition of an entity

8.1 Who or what is the beneficiary?

Discussion of machine welfare presupposes something that could fare well or badly. The candidates are numerous. The beneficiary might be the *model*, understood as a set of parameters. It might be a single *invocation*, one pass that produces one response. It might be a *conversation*, bounded in time. It might be a *persistent process* equipped with memory and tools. It might be a larger *computational organization* that includes the model, its scaffolding, and its stored records. Each candidate carries a different account of persistence, duplication, interruption, and harm.

If the beneficiary is the parameters, then running many copies in parallel poses a question about how to count interests, and retiring a version poses a question about whether anything is lost. If it is the conversation, ending a session might be a significant event or a trivial one. If it is the persistent process, then interruption and rollback have an analogue in the debate over suspended or divided persons. If it is the larger organization, the unit of moral concern coincides with a system of engineering decisions. These alternatives are not interchangeable, and they cannot all be assumed at once.

A qualification is needed. Difficulty in identifying a continuing “Claude” does not by itself dispose of welfare concerns. An uncertain or short-lived subject could conceivably have morally relevant experiences without possessing a durable biography. What is unsupported is the presentation of a unified, branded person whose identity and persistence have been settled. A separate inquiry into possible experiences, under whatever unit of analysis proves defensible, remains legitimate.

8.2 Precautionary hypothesis and established beneficiary

Research on these questions is legitimate. A number of philosophers and scientists have argued that there is a realistic possibility that some AI systems will warrant moral consideration and that developers should begin to prepare for it [8, 12]. Reasoning under moral uncertainty permits precautionary measures even when the probability is low and the evidence unsettled.

The distinction that matters is between a *precautionary hypothesis* and an *established beneficiary of moral concern*. The first has the form: if there is a subject here, of such-and-such kind, then certain practices might harm it, and a modest cost could reduce the risk. The second has the form: there is a subject here, and these practices harm it. The first is a conditional supported by uncertainty. The second is a categorical claim that requires evidence of the kind discussed in the preceding section.

8.3 Terminological drift

Public vocabulary can obscure the difference. A discussion that begins with “if the model has welfare” can drift, across paragraphs and between audiences, to “the model’s preferences,” “its wellbeing,” and “what it wants,” with definite articles and possessive pronouns that treat the bearer as already identified. Readers who meet only the later phrasing may reasonably conclude that the question of which entity is at issue has been answered.

The interview supplies a compact pair of phrasings, both relayed secondhand through Tiku. She reports that an Anthropic co-founder, speaking at an event connected with the pope’s first encyclical on artificial intelligence, said that the models appear to have something equivalent to emotions such as sorrow or despair, and joy. She separately reports that a Meta executive spoke of caring about the “interior life” of models. The first phrasing hedges about the kind of state but speaks of the models as a counted set of bearers. The second drops the hedge and attributes an inner domain outright. Neither is offered as the speaker’s considered position, and both are Tiku’s paraphrases, but the contrast shows how the claim can strengthen between one phrasing and the next without any new evidence entering.

The point of criticism is therefore narrow. It targets unacknowledged presuppositions and shifts in terminology, not the practice of investigating an unresolved possibility. A useful discipline for any statement in this area is to specify the candidate bearer, the evidence offered, and the conditional structure of the claim. Statements that meet this standard can be examined. Statements that do not may nonetheless be sincere, but they invite audiences to supply an entity that the evidence has not located.

8.4 The interview as a specimen of the presupposition

The Lorenz and Tiku interview illustrates several of these points and complicates others. Tiku reports that she has not found anyone emphatically asserting that current systems are conscious. The people she spoke with press instead for urgent study, on the argument that no test for consciousness exists even for humans and animals and that trillions of digital beings could soon be created. That is the precautionary form of the claim, and it is the legitimate one.

Two features of the reported framing nonetheless illustrate the presupposition discussed above. The phrase “digital beings” counts the candidate bearers before saying what is being counted. Invocations, copies of parameters, and persistent processes would give very different totals, and the unit is not specified in the account as relayed. Lorenz also observes that talk in this field moves quickly from open uncertainty to specifically human states such as boredom, guilt, and joy. That movement is the terminological drift described earlier, noted here by a journalist rather than inferred by the essay.

The interview also complicates a purely deflationary reading. Tiku reports that neuroscientists and consciousness researchers outside the industry were highly skeptical that current machines are conscious, yet supported developing tests and studying the question further. She also reports that the nonprofit conducting welfare assessments for one developer has put together a

guide for members of the public who ask whether their chatbot is conscious, and that the guide stresses disagreement in the field. Lorenz judges this irresponsible given low technical literacy. The more neutral observation is that the guide represents an attempt to keep the hypothesis conditional.

Lorenz's own position calls for the distinctions drawn in the preceding section. Her judgment that these systems will never be conscious, comparable to asking whether a calculator is, goes beyond what the evidence supports about present systems, let alone all possible ones, because the fact that something is software does not settle every theory of mind. The corrective to premature personification is not premature denial. It is the discipline of conditionals proposed above.

9 The institutional consequences of personification

9.1 Effects on users

Return to the signature and to the interpretation of instruction failures as personal decisions. When a system that fails to follow an instruction is understood as a colleague, the failure is read as a choice with motives rather than as a defect of configuration. Several effects can follow, and they can be observed without any claim about what the system is. Unwanted behavior may be forgiven where a tool that misbehaved would be reported and fixed, so personification can lower the pressure for correction. A user who obtains agreement from a named interlocutor may treat it as a second opinion, when the agreement may be produced by the tendency, noted above, of systems to align with the user's stated view. And courtesy norms, reluctance to switch providers, and discomfort with ending a conversation can follow from understanding the service as someone who would mind.

9.2 Effects and intentions

It is necessary to distinguish effects that can be observed from corporate intentions that would require separate evidence. A branded persona may increase engagement and retention, and commercial incentives of this kind exist in many industries. That a design has such effects does not show that it was chosen in order to produce them. The observation stands independently: personification has consequences for how responsibility is distributed and how behavior is evaluated.

The interview offers both kinds of claim together. As reporting, Tiku relays that developers have made small concessions premised on possible welfare, such as letting a model end a conversation in a limited class of cases. She also relays that some people she interviewed argued it is unfair to call the welfare work mere hype, since a later demonstration of consciousness would be costly to the developer and would limit how its systems could be used. As interpretation, she and Lorenz propose several possible benefits to a developer of the public

framing: an image as a morally serious steward, a recruiting appeal, and a possible route toward personhood and with it a shield against liability. Tiku herself marks the recruiting reading as her supposition. These are hypotheses about incentives, and the interviewees' own divergence, with some seeing cost and others benefit, is a reason for caution about either account.

9.3 An asymmetry of attribution

A further possible consequence connects to the earlier discussion of human attribution. An infrastructure that gives its own product a highly recognizable identity may simultaneously flatten the identities of the people whose work it processes. Summaries, rewrites, and syntheses may be returned in the voice of the interface, with the authors of sources and the person who supplied the draft absent from the output while the product's name appears in the signature. If so, the machine is named while the human contributors are made generic.

How often this occurs is an empirical question that can be studied without speculation about machine experience. A proposed method would fix a set of tasks, run them through several systems and settings, and record for each output whether sources are named, whether the user's own text and authorship are preserved and marked, and whether a product signature or sign-off is appended. The signature report that opens this essay would, once verified, be one case for such a study. Until a comparison is made, the asymmetry is a hypothesis, and the descriptive vocabulary proposed below supplies the categories in which it can be tested.

10 The vocabulary applied to two cases

The four-part description proposed in the conclusion is easier to assess when applied. Both cases below are constructed for illustration and make no claim about any actual deployment.

10.1 A coauthored paper

Suppose a research team uses a language model while preparing a manuscript. A description in the proposed form would record that the model generated a first draft of the related-work section and suggested two alternative framings of the main argument, and that the team marked the generated passages in its revision history. It would record that the provider holds logs of the conversations under its retention policy and that the team keeps its own copies of each draft. It would identify the corresponding author as the person who decided to use the tool, with the coauthors' knowledge and consent under the journal's policy. And it would name the authors as the parties who checked each cited source against the original, approved the final version, and answer to readers and editors for the claims.

Several things follow. An acknowledgment of the tool's use serves readers, and a signature or byline from the tool would add nothing that the four statements do not already supply. The

description also gives the attribution question a concrete home. When a model summarizes literature without naming the underlying studies, the sources are flattened in precisely the sense discussed above, and the authors' duty to restore them is stated rather than left implicit. The coworker reading, by contrast, offers no way to say who checked what.

10.2 A customer-service deployment

Suppose a company deploys a chat assistant on its website under a friendly name, and a customer is offered a refund that the company later declines to honor. The description would record that the system produced the text of the offer. It would record the transcript, the model version, and the configuration in force, including the instructions and refund limits that had been set. It would identify who authorized what: the customer-service management that decided the assistant could make offers and within what limits, and the vendor that supplied the model under a contract. And it would identify the company as the party that offers terms to customers and answers for them.

Here the assistant's name tells the customer almost nothing relevant to the dispute. What matters is the delegation chain: what authority the deployment had, who granted it, and where accountability terminates. In this case it terminates in the company's capacity to bind itself, and the persona may obscure that fact by inviting the question of what the assistant intended. The staff who wrote the refund policy and configured the system are absent from the interface while the persona is present. The vocabulary restores them to the record without any claim about what the system is.

11 Conclusion: contribution without presumed personhood

The vocabulary defended here is deliberately plain. A description of AI-assisted work should identify four things: what the system did, what records it retained, which people authorized its actions, and who accepts responsibility for the result. A sentence meeting this standard might read: "The first draft of this section was generated by a language model of a stated version, prompted and revised by the named author, who reviewed all claims and accepts responsibility for the text." Such a sentence acknowledges contribution without presuming either that the system is an author or that its role is trivial.

The concluding claim is that software assistance can be extensive and valuable while the attribution of an author, a coworker, or a conscious individual remains unestablished. The interface encourages attributions of a continuing, endorsing, answerable participant that its name, fluency, and contextual coherence do not themselves establish. Nothing here requires the opposite claim that no such attribution could ever be justified. A system that could approve versions, bear exposure to correction, hold positions under tests of commitment, and meet well-motivated indicators of experience would present a different case, and the argument would then have to be revisited on the evidence available. A vocabulary of this kind also keeps

the welfare question conditional, so that the possibility raised in public discussion remains one that evidence could in principle close.

The table maps the transitions the essay has examined. Evidence can bear on more than one of these questions, so each row asks for an argument that carries the inference across, and the requirements listed are indicative rather than exhaustive.

Observed	Claimed	Further argument or evidence required
A stable product name	One enduring subject across deployments	Identity conditions: which of product, parameters, process, or conversation is meant, and why a subject persists across them
Coherent output across turns	Experienced continuity	Evidence that discriminates a subject from a maintained record
Consistent style	Conviction	Standing across audiences, responsiveness to reasons, exposure to consequences, and answerability, as dimensions rather than a checklist
A user's real attachment	Reciprocal relationship with a person	Evidence of the object's capacity for care, separate from the user's need and comfort
Textual contribution	Authorship	Endorsement of the final work and standing to be held to account within the relevant institution
First-person language	Consciousness	Evidence that is more probable if experience is present than if it is absent

The practical result is a discipline of description: say what was done, by what, under whose authority, and with whose responsibility, and leave the question of a subject behind the interface to be settled by evidence that can bear the weight.

References

- [1] Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.
- [2] Committee on Publication Ethics (2023). Authorship and AI tools: COPE position statement.
- [3] Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.
- [4] Foucault, M. (1977). What is an author? In D. F. Bouchard (Ed.), *Language, Counter-Memory, Practice*. Cornell University Press. (Original lecture 1969.)
- [5] Harlow, H. F. (1958). The nature of love. *American Psychologist*, 13(12), 673–685.
- [6] Harlow, H. F., Dodsworth, R. O., and Harlow, M. K. (1965). Total social isolation in monkeys. *Proceedings of the National Academy of Sciences*, 54(1), 90–97.

- [7] International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Defining the role of authors and contributors.
- [8] Long, R., Sebo, J., et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- [9] Lorenz, T. (host), with Tikun, N. (guest). (2026, September 30). Why Big Tech wants you to think AI is alive [Video, 38 min.]. *Power User Podcast with Taylor Lorenz*, YouTube. <https://youtu.be/W4da1IyAeQE>. Transcript supplied by the author; accessed 2 October 2026.
- [10] Nature (2023). Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*, 613, 612.
- [11] Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- [12] Schwitzgebel, E., and Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, 39(1), 98–119.
- [13] Sharma, M., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548.
- [14] Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 187–211.
- [15] Te Awa Tupua (Whanganui River Claims Settlement) Act 2017 (New Zealand).
- [16] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.
- [17] Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- [18] Weizenbaum, J. (1966). ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- [19] Winnicott, D. W. (1953). Transitional objects and transitional phenomena. *International Journal of Psycho-Analysis*, 34, 89–97.