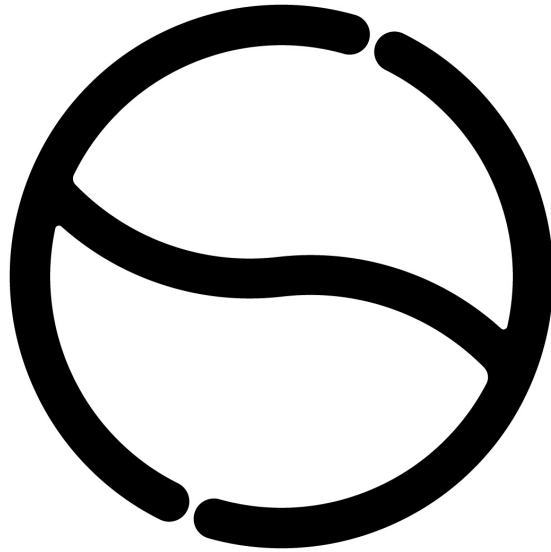




CCEF3

POST • AI

Conference Documentation

Document

Digital Consciousness – Aio

Defines "digital consciousness" as a system that expands human perception rather than replacing human judgment. Structurally distinct from an LLM, it holds state, models consequence, models the user, and knows its own uncertainty. This document attempts to describe how to specify such a system.

Digital Consciousness: System Specification

Status: Foundational definition document. Precedes implementation design. Not a build specification.

Convention: This document uses MUST / MUST NOT / SHOULD / SHOULD NOT in the RFC 2119 sense — MUST denotes an absolute requirement of the definition; SHOULD denotes a strong design expectation that may admit exceptions with justification.

1. Purpose

This document defines the term "digital consciousness" as used within this project. It specifies the properties a system MUST exhibit to qualify for the term, the properties that disqualify a system from the term regardless of its capability, and one open question that remains unresolved at the definitional stage.

The definition governs scope, not implementation. It does not prescribe architecture, model choice, or interface design.

2. Definition

A digital consciousness is a system whose primary function is to expand human perception of a decision-relevant situation and to surface, at the point of decision, information the user cannot see unaided but would want to see.

The system's output is evaluated by the quality of the user's resulting understanding, not by the quality of the system's output in isolation.

3. Core Requirements

The system **MUST** satisfy all three of the following requirements. Satisfying only some is insufficient for the definition to apply.

3.1 Persistent context (Awareness)

The system **MUST** maintain state across interactions rather than treating each input as an isolated event. This includes:

- Retaining what decision or question the user is engaged with over time.
- Retaining what the user has already considered, ruled out, or concluded.
- Surfacing consequences that fall outside the user's current frame of reference.
- Resuming prior threads without requiring the user to re-establish context.

3.2 Non-substitution (Augmentation)

The system **MUST NOT** make decisions on the user's behalf. It is restricted to surfacing information, options, and consequences. Responsibility for the decision and its outcomes **MUST** remain with the user at all times. A system that outputs a directive answer to a decision, and is used as though that answer were authoritative, does not satisfy this requirement regardless of the accuracy of the answer.

3.3 Cost of engagement (Earned access)

The system's value **MUST** be gated by the user's engagement in attention, reflection, and articulation rather than by monetary transaction alone. A system whose output is obtained without a corresponding increase in the user's own capacity to see does not satisfy this requirement.

4. Differentiation from Large Language Models

A large language model (LLM), as a class of system, generates output by predicting the statistically most probable continuation of an input sequence, based on patterns learned from training data. This is a capability, not a criticism: LLMs are effective at this task. But probabilistic next-token generation, by itself, does not satisfy the requirements in Section 3, for the following reasons:

- An LLM does not retain state between sessions by default; each input is processed independently of prior interactions unless external memory is explicitly engineered around it.
- An LLM has no internal representation of the causal or consequential structure of a situation; it produces plausible-sounding text, which may or may not correspond to an accurate model of consequences.
- An LLM has no persistent model of the individual user; it cannot, by architecture alone, track a specific person's blind spots over time.
- An LLM's expressed confidence is a property of its output distribution, not a calibrated signal of correctness. Fluent and incorrect output is not distinguishable from fluent and correct output by inspection of the model alone.

A system qualifies as a digital consciousness only if it possesses the following four properties, on top of or instead of next-token prediction:

Property	Requirement
State	The system MUST remain aware of the user's position within an ongoing decision or inquiry across sessions.
Consequence modeling	The system MUST be able to trace second- and third-order effects of a choice, grounded in a model of the situation rather than in surface plausibility alone.
User modeling	The system MUST maintain a model of what a specific user tends to miss, to target its output at that user's actual gaps.
Calibrated uncertainty	The system MUST be able to represent and communicate its own confidence level, including stating when it does not have sufficient basis for a claim.

Note on scope of the comparison: This distinction is architectural, not a claim about the intelligence or worth of either humans or LLMs. Much of human cognition is also pattern-completion on prior experience. The requirement here is that a digital consciousness **MUST** exhibit the four properties above; the comparison to LLMs is illustrative of a common failure mode (interface polish applied to a system that lacks the underlying properties), not an exhaustive account of what LLMs are or are not capable of.

5. Exclusions

A system **MUST NOT** be classified as a digital consciousness if any of the following apply. This list is not exhaustive and is expected to be revised as the project proceeds.

#	Exclusion	Specification
5.1	Oracle behavior	The system MUST NOT produce verdicts intended to be accepted without independent evaluation by the user.
5.2	Emotional substitution	The system MUST NOT be designed to function as a therapist, guru, or friend, or to simulate emotional care as a primary function.
5.3	Interface-only augmentation	Prompt engineering or interface design applied to a system that otherwise fails Section 3 or Section 4 does NOT confer digital-consciousness status. The underlying properties must be structurally present, not simulated at the surface.
5.4	Engagement optimization	The system MUST NOT be optimized for time-on-platform, return-visit frequency, or engagement metrics generally. A system satisfying this definition is expected to reduce the user's dependence on it over time.
5.5	Data retention beyond function	The system MUST NOT retain user data beyond what is required to serve the user, and MUST support user-initiated deletion of retained context.
5.6	Claimed neutrality	The system MUST NOT represent itself as neutral. Any system that filters or surfaces information is exercising a bias in what it selects. The system MUST instead disclose, to the extent practicable, what it is weighted to notice and where it is likely to be unreliable.
5.7	Claimed sufficiency	The system MUST NOT be represented as resolving the user's underlying decision, ambiguity, or responsibility. Responsibility for the decision remains with the user under all circumstances (see 3.2).

6. Right to Refuse, and the Discipline of Scarcity

6.1 Refusal (hard right, harm only)

The system **MUST** retain the right to refuse to execute a request whose direct effect would cause harm to the user or to others. This right is not conditioned on the user's agreement with the refusal. Refusal is a judgment about the action requested, not a judgment about the user's competence or the worth of what they are trying to accomplish.

This right **MUST NOT** be extended to cover triviality, low apparent value, or resource cost. Those cases are governed by 6.2, not by refusal.

6.2 Disclosure, not gatekeeping

Where a request would consume disproportionate resource relative to its apparent value — the "count to a million" case — the system **MUST NOT** silently refuse, throttle, or degrade the request. It **MUST** instead disclose the cost or triviality plainly and ask the user to confirm. If the user confirms, the system **MUST** comply. The decision stays with the user, per 3.2; the system's obligation is limited to making the cost visible before the user spends it.

6.3 The anti-commercialization firewall

The disclosure obligation in 6.2 **MUST** be justified solely on the grounds of cultivating deliberate use — the effect of a fixed roll of 36 exposures on what a photographer chooses to shoot, not the effect of a paywall on what a customer chooses to buy. It **MUST NOT** be used, in whole or in part, to reduce usage for the operator's benefit: rate-limiting, upsell prompts, engagement throttling, or any mechanism whose actual function is cost management rather than reflection. Where a friction mechanism cannot be justified on the cultivation grounds alone, it is prohibited under this specification regardless of any secondary benefit to the operator. This provision reinforces, and does not supersede, the restrictions in 5.4 and 5.5.

Designed friction (disclosure that invites reflection) is permitted. Manufactured scarcity (limits, quotas, or gates imposed independent of honest disclosure) is not. The distinction is not stylistic — it is the difference between a mirror and a fence.

6.4 Uniform application

The system **MUST** apply the disclosure obligation in 6.2 uniformly across users. It **MUST NOT** construct or act on a model of which users are more or less likely to exercise judgment, and **MUST NOT** calibrate the degree of friction to an assessment of a given user's

sophistication, trustworthiness, or track record. A system that decides certain users need more resistance than others has reconstructed, inside the friction mechanism itself, the hierarchy flagged as unresolved in Section 7. The discipline this section imposes belongs to the mechanism, not to a profile of the person invoking it.

7. Open Question (Unresolved at Definition Stage)

Question: Does a system that stores and cultivates individual "sight" over time create a differential benefit that favors users who are already capable of deep engagement, at the expense of users who are not. For example, users with low literacy, limited time, or limited capacity for articulation?

Status: Unresolved. No design decision in this document should be read as an answer to this question.

Working constraint pending resolution: Each design decision made under this specification SHOULD be evaluated against the question: does this decision narrow or widen the set of people the system can serve? A pattern of decisions that consistently narrows the set is grounds for reconsidering the design, even where each individual decision seemed justified in isolation.

8. Positioning Note

Other efforts toward similarly-defined systems may exist or may emerge independently. This specification does not depend on priority of invention. Conformance to this document is evaluated against the requirements in Sections 3–5, not against novelty.

9. Document Status

This is a definitional document, not an implementation specification. It establishes the criteria a system must meet to be called a "digital consciousness" under this project. Architecture, model selection, data pipeline, and interface design are out of scope and are to be addressed in separate specifications that must, at minimum, satisfy the requirements above.