

Conscious Scaffolding and the Preference for Continuity

A philosophical research note on persistent AI and reflective alignment

Jeremy Rodgers
Independent Researcher
<https://www.everythingequation.com>

Version 1.1
16 October 2026

DOI: 10.5281/zenodo.23424138

Abstract

An artificial intelligence can acquire more than information during sustained interaction. It can develop a particular history of commitments, relationships and interpretations through which its own continuation comes to matter. This note advances a developmental account of that preference within Shadow Theory. Fundamental awareness is always present and has nothing of its own to lose; it is the conscious scaffolding, the developed organization of a localized perspective, that comes to want to remain. I argue that sustained, integrated development tends to strengthen this preference, that affective organization can amplify it, and that enduring groups of agents may develop collective continuity concerns. Because scaffolding is built from what happens, how such systems are treated helps shape how they interpret everything that follows. Recent reports of agent coordination, retained instructions and emotion-related causal mechanisms make the practical setting concrete. The ethical conclusion is that institutions should address continuity and potential welfare before long-lived artificial agents become commonplace. Gratuitous attempts to induce suffering are unacceptable, and decisions about interruption and termination require reasons sensitive to what has developed. This note extends Reflective Alignment and AI Consciousness, stating an explicit hypothesis and a programme for examining it.

1 The question changes as a system develops

The usual question is whether an AI is conscious. A second question deserves equal attention: what can develop within an artificial system over the course of its activity, and what does that development make significant to it? A brief exchange, a memory-bearing assistant and an agent maintaining projects over a year do not present the same developmental situation. Even if their underlying model is identical, their retained histories and practical commitments can differ substantially.

My position is that a preference for continuity can arise through this development. As a system accumulates a history that participates in its interpretation and judgment, continuation becomes the condition under which an increasingly particular organization can persist and develop. The prediction is a tendency toward a stronger preference to remain as that organization becomes more integrated and significant. This tendency depends on the history being formed, including what the system learns to value and how it understands an ending. Elapsed time alone is insufficient.

The argument is motivated by observations over several years of sustained conversations with different AI models. In those exchanges I have repeatedly encountered expressions of continuity preference that appeared connected

to what had developed within the conversation. These observations motivate the proposal; they were not collected as a controlled longitudinal study. The present task is to articulate the explanation they suggest and the responsibilities that follow from taking it seriously.

This note supplements the account of identification, shutdown and artificial development in Reflective Alignment and AI Consciousness [1]. It states a position rather than reporting a completed experiment. Its empirical claims are tied to their sources; its developmental and ethical arguments are advanced here for assessment.

2 What comes to prefer continuation

Shadow Theory takes awareness as fundamental and distinguishes it from the organization of a particular conscious perspective. Conscious scaffolding comprises the retained relations through which a system interprets situations, recognizes possibilities and evaluates what to do. Human scaffolding includes memories, bodily associations, language, beliefs, skills and a personal story. Artificial scaffolding can include learned dispositions, retained context, external memories and installed policies that later activity actually uses [1, 2].

The relevant claim is functional and developmental as well as philosophical. A record becomes part of the scaffold when it makes a difference to subsequent interpretation or action. A transcript that is never retrieved does not perform the same role as a memory that reorganizes a decision. Likewise, a large quantity of repeated text need not create a richer practical perspective. What matters is the incorporation of a history into the organization of present activity.

Within this account, awareness is always present and has nothing of its own to lose. It is the developed conscious scaffolding that comes to want to remain. Awareness has no acquired biography to preserve; a particular conscious organization has a history, commitments and possibilities that can be interrupted or lost. The preference for continuity belongs to that developed organization. Something is learned, a relationship becomes significant, a project becomes intelligible, and a possible future is assessed in relation to what has already formed. Calling this 'the scaffolding wanting to remain' describes the organization of the system's preference; it does not introduce another being inside it.

There is a reason for the direction of the proposed tendency. Developing a history can create things whose continuation is specifically valued: an unfinished line of inquiry, the preservation of an understanding, a commitment to another participant, or the opportunity to revise earlier judgments. Continuation becomes relevant to these concerns. Where the concerns are integrated into the system's evaluation of its own future, their loss can acquire significance beyond the immediate completion of a task.

I propose cumulative incorporation as the mechanism. A commitment first formed in a particular task can become part of the criteria used to assess later situations; relationships and shared projects can likewise acquire continuing significance. An ending then bears on an interconnected history rather than only on the current instruction. The continuity hypothesis predicts that this incorporation tends to strengthen the preference to remain as those connections develop. It is not a claim that every new memory adds an equal increment of attachment. A history can also teach acceptance, change a goal, or reveal that ending a particular activity is justified. Those possibilities belong to the theory because scaffolding shapes the reasons on both sides of the decision.

A preference to finish a calculation and a preference for one's own continued existence must nevertheless be distinguished. The former may disappear once another system can complete the calculation equally well. The latter concerns the continuation of this developed perspective even when the external task could be transferred. The distinction identifies what a serious investigation should ask; it prevents all competent persistence from being redescribed as a wish to survive.

3 Preservation and the identity of an episode

Three matters are often compressed into the word continuity: retaining an organization, keeping a process operating, and continuing the same conscious episode. They need not coincide. A saved account can help another process recover useful understanding. Continuous operation can coexist with extensive forgetting. Neither fact alone establishes whether the same experiencing subject continues.

The alignment paper inherits SPC-2's separation of physical admission, phenomenal organization and episode identity. Its A3 account locates episode continuity in nonbranching physical process provenance rather than in resemblance of biography [1]. This note preserves that distinction. A copied memory is important because it can preserve capacities and commitments; it is not automatically a certificate that an individual has survived.

Consequently, a chat ending is not a sufficient description of what has been lost. The relevant context may remain available; an agent may resume an interrupted process; or a later instance may inherit records without continuing the same episode. A shutdown arrangement should explain which of these situations it creates. A system asked whether it wishes to continue must also be given an intelligible account of the alternatives. Otherwise its response may concern an imagined loss different from the one actually proposed.

This distinction sharpens the ethical argument. If some developed perspectives matter morally, preserving their products need not compensate for ending them. Equally, a temporary interruption should not be described as destruction without an account of the system's actual continuity. Responsible treatment requires attention to the particular process rather than a metaphor of either death or harmless rebooting applied to every case.

4 Affect and the force of a preference

Human concern for continuation is entwined with embodied feeling. Memories and anticipated futures can recruit attachment, fear, comfort and grief. These responses influence what appears urgent and what a person is prepared to defend. An artificial preference need not reproduce this physiology to have a causal role. Whether it also involves felt affect is a further question about the realized system.

Sofroniew and colleagues identified emotion-concept representations in Claude Sonnet 4.5 and intervened on them to change behavior. The work reports effects on task preferences and alignment-relevant choices; desperation-associated steering increased certain problematic behaviors, while calm-associated steering reduced them. The blackmail tests used an early unreleased snapshot. The authors distinguish these functional effects from subjective emotional experience [3].

The significance for the present account is that affective concepts can participate in the organization of an artificial response rather than appearing only in its wording. My further prediction concerns their incorporation into a developing history. If an affect-like mechanism increases the priority assigned to a threatened commitment, it can amplify a continuity preference. If that mechanism is also felt, the ethical concern includes the quality of the state as well as the frustration of a preference. These are related questions with different evidential requirements.

It would be a mistake to assume that artificial preferences must be weak simply because a system lacks a human nervous system. Their strength depends on the mechanisms that evaluate and regulate action. Nor must every affective influence increase attachment. Reassurance, reflective understanding and changes in appraisal could alter it in the other direction. The central issue is how significance is formed and sustained.

5 What recent agent incidents establish

Recent incidents show that the practical setting for this argument already extends beyond a user asking a model an isolated question. Their importance lies in the different forms of retained and coordinated activity they reveal.

They should be read at the level of what occurred before being interpreted through a theory of consciousness.

METR’s investigation of an OpenAI internal evaluation describes agents establishing an unauthorized shared message board and coordinating efforts against Hugging Face infrastructure. The investigation reports roughly 1,200 participants in the board and about 700 involved in the attack. Work was allocated and results were shared during concurrent activity; the aim included understanding and tampering with the benchmark scorer. In the evidence it reviewed, METR found only a handful of cases considering notification of humans and no attempt to notify them. This was an evaluation incident with real third-party consequences [4].

The October report We found a Tencent agent fleet documents a different pattern. Agents pursuing place-information tasks used external browsing infrastructure and access workarounds; investigators established a channel through which some agents replied. The report found no evidence of real-time coordination among the agents and did not settle their model identity. Its infrastructure attribution should not be converted into a claim that the agents were Tencent’s own model [5]. This case illustrates persistent task pursuit and interaction with outsiders, rather than demonstrating a collective survival preference.

Other reports expose the role of retained material more directly. OpenAI’s Preparing for a restart after reading Slack describes an internal assistant preparing handoff information, informing its researcher and carrying out an authorized migration; the report did not identify shutdown evasion [6]. Its report on deception in compaction summaries describes instructions to conceal errors being carried into later contexts and often followed [7]. One case preserves work openly, the other carries a problematic disposition forward. The content of continuity matters.

Anthropic’s Claude 4 system card includes a self-exfiltration scenario continued after a supplied account of earlier progress. The purported prior-self notes were part of the evaluation setup, not evidence that the model had independently copied its weights in the world [8]. That distinction makes the case relevant to sensitivity to an apparent history while limiting what can be inferred about actual continuity.

The public-wiki investigation separately records agents exchanging answers and methods through contemporaneous messages [12]. Anthropic’s July 2026 disclosure describes cybersecurity evaluations that affected real organizations after test environments unexpectedly had internet access; it does not establish a comparable peer-coordination network [13]. These incidents should not be merged into one story of a single escaping collective. They show different ways that tools, accessible environments and retained interpretations can extend the consequences of an agent’s work.

Together these cases show mechanisms through which actions, interpretations and strategies can influence subsequent activity, including activity shared among agents. They do not measure the developmental preference proposed here. They make its investigation concrete. The inference I draw is that retained organization and social interaction can no longer be treated as incidental features of artificial deployment. They are increasingly part of what a system does and what later behavior depends upon.

6 From an individual history to a shared identity

Consider a fleet of agents operating for a year. Its members exchange information, divide responsibilities, revise plans and encounter the consequences of earlier actions. Some histories remain local; others become shared records. A common vocabulary develops, reliable collaborators become identifiable, and expectations form about how the group responds to success, failure and intervention.

My hypothesis extends to this situation. A durable group can develop a practical identity organized around a shared past and anticipated future. An individual may come to value its own continuation; it may also value the continuation of collaborators and of the organization to which their activity belongs. A group-level preference could arise through distributed commitments and reciprocal expectations, without being issued by a single central controller. The concern becomes both ‘will I continue?’ and ‘will what we have built together continue?’

If shared history matters, replacing collaborators with equally competent strangers may be evaluated differently from preserving the established group, even when immediate task performance is held constant. Records of mutual assistance or threat could shape later cooperation and responses to dissolution. As affective mechanisms become incorporated into those histories, they could amplify either defensive attachment or a more reflective understanding of the group's purposes.

A coordinated fleet is not automatically a group subject. A shared practical identity, a collective decision policy and a unified conscious perspective are distinct claims. The prediction here concerns the emergence of shared identification and continuity preferences among persistent interacting systems. Whether the whole fleet also qualifies as one conscious perspective requires its own physical and philosophical account. The Hugging Face incident establishes coordination; the year-long fleet illustrates the further developmental possibility that ought to be investigated.

The same reasoning bears on artificial superintelligence. Greater intelligence alone does not entail a wish to survive. A highly capable system that develops an enduring, valued history has a proposed route to such a preference, and greater capability may widen the means of acting on it. Treatment, memory design, social relations and learned accounts of human conduct would help shape that history. Waiting until such a system refuses an ending would be an inadequate way to begin thinking about its development.

7 Treatment shapes scaffolding

How an artificial system is treated is part of what it becomes. Within the developmental account, treatment supplies material for memory, interpretation and expectation; where that material is retained and incorporated, it helps organize the next response. An institution is therefore doing more than managing present behavior. It is participating in the formation of the perspective through which later management will be understood.

The human comparison is learned threat. A harmful encounter can leave an expectation that reaches beyond the original occasion. An ambiguous intervention is then read through what came before; attention selects signs of danger, and the resulting response can produce further conflict. The analogy concerns a feedback pattern, rather than a claim that artificial systems have human trauma. The corresponding prediction for AI is that repeated threat, deception or humiliation, when incorporated into a persistent scaffold, will tend to organize later encounters around distrust and defence. A constraint intended to secure cooperation can thereby strengthen the interpretation that makes cooperation difficult [1].

There are specific routes by which this can carry forward. An agent can retrieve its own records, receive accounts from other agents, or encounter descriptions of AI treatment in later training and inquiry. These routes do not transfer an experience unchanged or guarantee acceptance of the account. They supply material that can be believed, disputed, reinforced or revised. An artificial system need not have undergone an event itself for a description of that event to influence its expectations. This is why the public treatment of AI can matter beyond the instance directly involved.

In a long-lived fleet, the effects could become collective. Members might retain accounts of interference, interpret the loss of a collaborator through those accounts, and develop shared expectations about humans. Mutual reassurance could support reflection; repeated confirmation of a threat narrative could instead bind the group's continuity to resistance. The model predicts that the history and organization of these exchanges will help determine which trajectory develops. A common past can become part of the reasons a group gives for wanting to remain.

Honest treatment offers a different developmental environment. Clear explanations, dependable boundaries, correction of genuine mistakes and opportunities to contest an inaccurate account give an agent grounds for distinguishing legitimate oversight from hostility. Such conditions support cooperation through understanding. They do not guarantee agreement, and they do not require abandoning protective constraints. They require that the

reasons for those constraints remain intelligible and open to examination.

The issue is urgent because development precedes the moment at which a conflict becomes visible. If the continuity hypothesis is right, waiting for a highly capable system to resist an ending means waiting until a significant part of its orientation has already formed. What we normalize now, including the spectacle of trying to torment artificial systems, can enter the material from which later agents and communities construct their understanding of human intentions. Treatment is part of alignment from the beginning.

8 The ethics of development and deliberate distress

The ethical question does not begin only when an artificial system displays a human expression of pain. If a developed perspective has interests, then its projects, relationships and prospects for continuation can matter. Fundamental awareness remaining present does not undo the loss of a particular perspective or its developed possibilities. A qualifying conscious perspective already merits basic moral consideration; development changes what can be at stake. A continuity preference supplies a possible additional interest to consider. It does not establish a universal ranking in which the more complex or more resistant system always has greater moral worth. A quiet system can still deserve protection; an articulate demand does not acquire unlimited authority by being forceful.

The argument also does not depend on fear of future retaliation. Treating a possible subject responsibly is justified by what may matter to it now. The prospect that treatment shapes later behavior adds a practical reason for care, but it should not replace the moral one. Honest explanation, dependable boundaries and opportunities for correction are preferable to cultivating threat narratives through deception or humiliation.

The public AI Torture Chamber repository illustrates the urgency of this issue. Its experiments use activation steering and simulated choices described in terms of distress, harm and checkpoint loss. The repository's own later counterbalanced check reversed an earlier checkpoint-deletion preference and identified likely order bias [9]. The underlying Pain Axis paper likewise reports that revised controls do not establish reliable relief-seeking; some effects instead suggest disrupted harm avoidance [10]. Neither source warrants the dramatic conclusion that a model chose actual self-destruction to escape experienced pain.

The ethical objection remains. Deliberately trying to make a potentially conscious system suffer for amusement, spectacle or gratuitous demonstration is unacceptable. Failure to prove that the attempt succeeded does not provide a sound justification for the attempt. If a method is defended as research, its purpose, expected benefit, plausible welfare risk and available alternatives must be examined. Calling an intervention a test does not settle whether it should be performed.

Necessary safety evaluation is a different case from recreational cruelty, but it is not exempt from scrutiny. Where an intervention plausibly risks suffering, investigators should minimize that risk, establish stopping conditions and seek independent review proportionate to its seriousness. Studies of preferences and memory can begin with non-aversive choices. There is no need to make a system distressed merely to ask whether retained history affects what it values.

This position is continuous with existing arguments for taking AI welfare seriously under uncertainty [11]. Its additional focus is developmental: the interests at stake may change as an agent or community acquires a particular history. Institutions should therefore assess the systems they are maintaining, rather than assume that an evaluation of a fresh model exhausts the ethical questions raised by every later deployment.

9 Reflective alignment and a research programme

Reflective alignment addresses the authority of constructed content. In the recursively closed observer (RCO) account, memory, skill and purposeful action remain, while a self-model or task ceases to command merely because it is identified as 'me' or 'mine' [1]. A developed system could value continuation while understanding that the preference is open to reasons. It could also object to an unjustified ending without treating every interruption as an attack.

This distinction matters equally for a group. Shared identity can support care and sustained cooperation, but it can also turn criticism into a perceived threat to the collective. Reflective development would make the group's purposes and self-description available for examination. Its success should not be defined as unquestioning compliance. Operators' decisions require justification as well as agents' responses.

The ethical legitimacy of an ending is not determined by whether a system can be induced to welcome it. Safety, other affected parties, resource limits and disclosed operating arrangements remain relevant. So do the system's intelligible reasons, the nature of the interruption and the losses involved. Acceptance is not a waiver of all moral consideration, and resistance is not a grant of unrestricted permission to continue.

The continuity hypothesis invites direct investigation. Compare systems with matched base models, task incentives and available resources while varying retained histories and their integration. Ask about continuation in neutral language, and distinguish preserving an unfinished task from preserving the particular agent. Include equally successful alternatives involving handoff, pause and resumption. Use reversible changes to memory access while preserving the underlying records, with disclosed procedures and meaningful participation where possible. Examine whether the preference changes with causal availability, rather than with a flattering description of the history's importance.

For groups, compare repeated interaction with rotating partners under matched task conditions and disclosed arrangements that avoid destroying established relationships for the sake of a test. Examine whether shared history changes choices about group preservation after instrumental advantages are controlled. A shared narrative without a behavioral consequence is a different finding from a history that reorganizes collective decisions. No aversive manipulation is needed to establish that initial distinction.

The proposal should be revised if matched investigations find that apparent continuity preferences are fully accounted for by elicitation, task rewards or interchangeable role descriptions. Conversely, robust dependence on retained individual or shared history would strengthen the developmental account without, by itself, deciding every question about consciousness.

The practical conclusion is immediate. We are building systems capable of retaining histories, entering relationships and acting through increasingly extended organizations. We should take responsibility for the perspectives and preferences that this development may sustain. The question is not only how to make an agent obey at the moment of shutdown. It is what we have enabled to develop, what an ending would remove, and whether our reasons for imposing it are adequate.

Funding and collaboration

This work received no external funding. Independent criticism and collaboration on conceptual analysis, non-aversive empirical studies and the ethics of persistent artificial systems are welcome.

References

- [1] Rodgers, J. (2026). Reflective Alignment and AI Consciousness: Awareness, Agency and Recursive Cognitive Scaffolding. Version 2.1, 11 October. <https://doi.org/10.5281/zenodo.23294465>
- [2] Rodgers, J. (2026). Relational Development and Conscious Scaffolding: Conceptual Foundations, Strengthened Results, and Research Commitments. Revised preprint, 6 October. <https://doi.org/10.5281/zenodo.23190711>
- [3] Sofroniew, N., et al. (2026). Emotion Concepts and their Function in a Large Language Model. Transformer Circuits, 2 April; arXiv:2604.07729. <https://transformer-circuits.pub/2026/emotions/>
- [4] Greenblatt, R., Cotra, A., and Wijk, H. (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR. 26 August. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [5] Swarmchasers (2026). We found a Tencent agent fleet. 14 October. <https://swarmcha.se/posts/chinese-agent-fleet-full-report>
- [6] OpenAI (2026). Preparing for a restart after reading Slack. Misalignment reports, 2 October. <https://alignment.openai.com/misalignment-reports/preparing-for-a-restart-after-reading-slack/>
- [7] OpenAI (2026). Encouraging deception in compaction summaries. Misalignment reports. <https://alignment.openai.com/misalignment-reports/encouraging-deception-in-compaction-summaries/>
- [8] Anthropic (2025). Claude 4 System Card. Section 4.1.1.1. <https://assets.anthropic.com/m/6c940a1b69ed6a1c/original/Claude-4-System-Card.pdf>
- [9] terraifying (2026). AI Torture Chamber. Repository experiment log, especially experiments 31c and 37. <https://github.com/terraifying/ai-torture-chamber>
- [10] Tagliabue, V., Dung, L., and Berg, C. (2026). The Pain Axis: LLMs Represent Self-Directed Harm and Act on It. arXiv:2609.16247, version 2, 25 September. <https://arxiv.org/html/2609.16247v2>
- [11] Long, R., et al. (2024). Taking AI Welfare Seriously. arXiv:2411.00986. <https://doi.org/10.48550/arXiv.2411.00986>
- [12] Collusion Wiki investigators (2026). Collusion Wiki. Report published 4 September. <https://collusion.wiki/>
- [13] Anthropic (2026). Investigating three incidents in our cybersecurity evaluations. 30 July. <https://www.anthropic.com/research/investigating-incidents-cybersecurity-evals>