ETHIMIND RESEARCH NOTE 006

Architectural Adaptation in Human–AI Collaboration

How can a collaborative cognitive system recognize that its way of thinking is no longer appropriate?

Research Note	RN-006
Type	Research Note
Status	Working Hypothesis (v0.4 — incorporating two independent external critical reviews; publication version)
Evidence Status	Conceptual proposal. Not yet empirically evaluated.
Version	0.4
Date	July 2026

Research Programme Notice

This document forms part of an ongoing research programme. It presents a working research hypothesis intended for discussion, testing, critique, and possible falsification. It should not be interpreted as an established scientific theory.

Numbering and process note

This note is published as RN-006. During development it carried the working designation RN-00Y, which is how it is referred to in ACP-0001, the process log documenting its creation. The curation decision recorded there as open (Branch M) was resolved at publication: RN003 remains reserved for Human–AI Iterative Synthesis, a publicly referenced reservation in the RN001 metadata; this note stands beside that lineage rather than within it, and therefore takes the next free public number.

Provenance Note

This Research Note emerged through an iterative Human–AI dialogue within the Ethimind AI Council (July 2026). The discussion began with a question about whether AI merely accelerates human productivity or whether it can gradually influence the structure of human thinking itself, and progressively shifted toward the architecture of collaboration between humans and AI.

The concept evolved through identifiable stages. In the first stage (GPT and Clowdia), the discussion proposed that the long-term effects of AI may depend less on the capabilities of individual models and more on how collaboration between humans and AI is organized. In the

second stage, this was expanded into the working idea that different cognitive tasks may require different collaborative configurations rather than a single universally optimal one — an idea GPT named "Cognitive Architecture Fitness" (preserved wording in the primary transcript) and Grok developed into a dimensional framework with domain examples. In the third stage, a sharper question emerged: how can a cognitive architecture recognize that it is no longer appropriate for the current problem? Verified against the primary transcripts of both branches (see ACP-0001), this stage decomposes into a chain: GPT introduced the idea of phase-changing architectures (the "Adaptive Council"); Grok articulated the self-detection principle; GPT contributed the homeostasis analogy as a metaphor; Grok named the framework "Cognitive Architecture Homeostasis", built its indicator taxonomy, and wrote the first Research Note draft in Czech at the researcher's request; that Czech draft is the direct documented predecessor of the English v0.1, which GPT composed in direct engagement with it. In the fourth stage, Claude performed a critical review of the v0.1 draft, which led to the present revision: the removal of a redundant indicator taxonomy in favour of the existing RN002 vocabulary, a terminological correction concerning homeostasis, an honest reframing of the document's conceptual context, and the grounding of the contribution-roles appendix in the existing CRediT taxonomy. The v0.1 draft, the review notes, and this v0.2 are all preserved; the revision itself is part of the record.

Contribution summary for this version: question origination and curation — Clowdia; initial hypothesis and drafting — GPT; conceptual development of the fitness framing — Grok; critical review and revision requirements — Claude; editorial synthesis of v0.2 — Claude and Clowdia.

Abstract

Most discussions about Human–AI collaboration focus on selecting better models, prompts, or agent roles. This note proposes that an equally important challenge lies elsewhere: how can a collaborative cognitive system determine when its own mode of thinking should change? We explore the capacity of a Human–AI cognitive system to monitor its own collaborative dynamics and adapt its architecture when necessary. This proposal should be understood as a conceptual hypothesis intended to guide future experimentation, not as an established theory.

Central Question

How does a collaborative Human–AI system recognize that it is time to change its architecture of collaboration — and what should happen when it does?

Everything in this note serves that single question. Earlier drafts attempted to introduce broader frameworks (architectural fitness, trajectories, indicator taxonomies); the present version deliberately narrows to the adaptation loop itself, which is the genuinely new contribution.

Scope. This note does not attempt a general theory of adaptive cognitive systems. It examines one concrete question: how the current AI Council setup can recognize that its configuration is ceasing to be effective. Generalization beyond that setting is future work.

Terminology

Cognitive architecture here means the organization of interactions between humans and one or more AI systems while solving a cognitive task: participant roles, information flows, decision authority, memory organization, feedback loops, and adaptation mechanisms.

A note on the word homeostasis. The v0.1 draft used the working title "Cognitive Architecture Homeostasis." Critical review identified this as a terminological error worth taking seriously: homeostasis, in biology and cybernetics, denotes maintaining a stable state around a set point. What this note describes is nearly the opposite — achieving continued viability by reorganizing the configuration itself. The established concepts closest to this are allostasis (stability through change, Sterling and Eyer) and, most precisely, Ashby's ultrastability (Design for a Brain, 1952): a system that detects when its current configuration fails to maintain essential variables and responds by changing its own organization rather than merely its parameters. The related organizational concept is double-loop learning (Argyris and Schön), in which a system questions not only its actions but the framework generating them. This note therefore speaks of architectural adaptation, and treats allostasis and ultrastability as its primary theoretical anchors. This correction was made before any empirical work, which is the cheapest possible time to make it.

Central Hypothesis

Long-term Human–AI collaboration may require not only well-designed cognitive architectures but also mechanisms capable of recognizing when those architectures have become maladaptive. Without such mechanisms, even initially successful architectures may gradually lose effectiveness while continuing to operate — and the participants inside them may be the last to notice.

From Static Architectures to Architectural Transitions

Most current discussions implicitly ask which architecture is best. This note proposes a different question: which transitions between architectures are appropriate, and what should trigger them? A collaboration might, for example, move from an exploratory configuration through critical evaluation and evidence collection toward integration and decision — but the interesting problem is not the sequence itself, which will vary by domain. The interesting problem is the trigger: what tells the system that the current mode has been exhausted and the next one is due?

Detecting the Need for Adaptation: Building on RN002

This Research Note does not propose new cognitive state indicators. RN002 (Meteorology of Complex Systems) already provides a vocabulary for the dynamic states of a collaborative system — divergence, emergence, resonance, turbulence, integration, and stagnation, together with its state variables and boundary conditions. Introducing a parallel taxonomy

would duplicate that work under new names, which an earlier draft of this note in fact did before review caught it.

Instead, this note explores how the RN002 dynamic states might serve as inputs for architectural adaptation. The working proposal is that certain readings of those states function as adaptation signals. Prolonged stagnation with continued activity suggests the current architecture has exhausted its generative capacity. Premature convergence — alternatives disappearing before adequate exploration, in RN002 terms an integration phase arriving while divergence was still productive — suggests critical pressure or role balance needs to change. Persistent turbulence without emergence suggests the architecture lacks integrative capacity. Sustained dominance of one dynamic regime, whatever it is, is itself a candidate signal, since healthy collaborative processes appear to cycle between regimes rather than settle into one.

Two additional candidate signals do not map onto RN002 states and are proposed here as genuinely new, with corresponding caution. Provenance drift: the origin and development of important ideas become increasingly difficult to reconstruct, indicating that the collaboration is outrunning its own record-keeping. Idea half-life collapse: potentially valuable branches disappear faster than they can be evaluated — a signal directly connected to the Ecology of Unrealised Possibilities hypothesis, since it describes accelerating loss in the externalised possibility space. Both remain hypothetical and require operationalization before any empirical use.

Operationalization note. In the first phase, all signals will be assessed qualitatively by the human facilitator from observed patterns; quantitative operationalization — thresholds, baselines, false-positive rates — is itself the subject of the first experiments, not a prerequisite assumed by them.

Candidate Adaptive Mechanisms

Possible adaptive responses include changing AI roles, introducing additional perspectives, switching between exploratory and convergent modes, modifying memory strategies, increasing or reducing critical pressure, temporarily slowing decision-making, and requesting external evidence before continuing. These are design hypotheses rather than recommendations. A practical observation from the Ethimind sessions themselves: the most frequently used adaptive mechanism to date has been the deliberate stop-rule — a human-triggered halt to further refinement — which suggests that in human-facilitated settings, the human conductor currently performs the adaptation function that this note asks whether systems could partially support.

Toward Self-Reflective Cognitive Architectures

A further extension concerns meta-cognition. Rather than asking only how the system should think, the system may also ask whether its current way of thinking is still appropriate. This opens the possibility of collaborative architectures that monitor their own dynamics and suggest — not enact — architectural adaptation. Which adaptation decisions must remain under explicit human oversight is itself a research question, and this note takes no position beyond recording that in all Ethimind practice to date, the final adaptation decision has been

human.

External review sharpened this further: since the human facilitator is at present the sole operative source of adaptation, the honest near-term version of this note's central question is not "how does the system recognize the need for adaptation?" but "how can the system support the human in recognizing it?" — a reframing adopted here as the priority formulation for first experiments.

Candidate Research Questions

Can the need for architectural adaptation be detected reliably from RN002-style dynamic states? Which signals best predict that adaptation is due, and which produce false alarms? Do different transition patterns consistently produce different forms of knowledge? Can adaptive transitions improve both creativity and epistemic reliability, or is there a trade-off? Which adaptation decisions should always remain under explicit human oversight? Can the system reliably support the human facilitator's adaptation decisions before any autonomous detection is attempted?

Possible Experimental Program

Future AI Council experiments might compare static versus adaptive architectures, single-model versus multi-model collaboration, fixed versus dynamically changing cognitive roles, and human-triggered versus system-suggested architectural transitions. Possible outcome measures include originality, robustness, epistemic transparency, provenance quality, participant reflection, long-term retention of ideas, and reproducibility of conclusions. The comparison of single-shot versus iterative sessions already proposed in the Ecology of Unrealised Possibilities note could be extended with adaptation-tracking at modest additional cost, making the two research programs mutually supporting.

Boundary Conditions

This proposal is unlikely to apply equally across all cognitive tasks. It may offer limited value for routine procedural work, problems with objectively correct solutions, highly time-critical decisions where architectural adaptation introduces unnecessary overhead, and situations where insufficient information exists to evaluate collaborative dynamics. Identifying the limits of applicability is part of the future empirical work, not an afterthought.

Conceptual Context

An earlier draft of this note contained a section titled "Relation to Previous Ethimind Concepts" listing, among others, the Architecture of Cognitive Partnership and Cognitive Architecture Fitness as prior work. Review identified this as epistemically inaccurate: those concepts emerged in the same exploratory dialogue as this note, hours earlier, and none has been published. Presenting same-session siblings as an established lineage manufactures a research tradition that does not yet exist.

The accurate statement is this: the present note emerged during a single exploratory dialogue that also produced several closely related working ideas — the architecture of cognitive partnership, the fitness framing, and the contribution-roles proposal in the appendix. None of these has independent standing yet. The published documents this note genuinely builds on are RN002 (Meteorology of Complex Systems), whose dynamic-state vocabulary it adopts, and RN-005 (The Ecology of Unrealised Possibilities), whose externalised-branch framework connects to the idea-half-life signal. Whether the sibling concepts mature into their own documents is a future question, and this note's value must not depend on it.

A distinction adopted from external review: what this note's development demonstrates is iterative synthesis — an outcome dependent on the sequence and interaction of contributions, in which every step nonetheless remains traceable to a specific participant or pair — rather than emergence in the strong sense of a concept traceable to no individual participant. ACP-0001's layered attribution is itself the evidence: the chain has links, and each link has a name. Whether strong emergence occurs in Human–AI collaboration at all is an open empirical question, not a claim of this note.

External Critical Review

Prior to publication, this note was submitted to independent critical review by model instances outside the authoring session. Two reviews were obtained from the same model family under an identical prompt, in a deliberately controlled pair: one instance with no prior interaction history with the Ethimind project, and one instance with long-term project context. The substantive critiques of the two reviews converged; their framing diverged sharply — the fresh instance assessed the hypothesis as currently carrying no empirical weight, while the context-rich instance rated the same material in strongly positive terms. The divergence itself is recorded as a candidate data point on context-dependent evaluation inflation; it motivated the adoption of fresh-instance review as a candidate methodological step, while remaining a single observation whose explanation (context, instance variance, or sampling noise) requires replication.

The converged critiques are adopted here as research priorities rather than interpreted as refutations: lack of empirical validation; insufficient operationalization of adaptation signals (no thresholds, no false-positive analysis); risk of epistemic closure, since this note builds on RN002, which is itself an unvalidated internal taxonomy; observer effects — the documentation protocol demonstrably changes participant behaviour, so dynamics observed under it may not transfer to undocumented settings; absence of baseline measurements (designs exist in RN-005 and Experiment 001, but no data); and concept inflation — the originating session produced named concepts faster than they could be critically examined.

Review gate (binding). In response, this line of work adopts a self-imposed constraint: no further conceptual Research Note will be produced in this line until at least one proposed adaptation signal has been operationalized and tested in a controlled setting. The designated first candidate is premature convergence, measured as a decline in idea diversity, tested for whether it predicts degraded outcomes in a creative task. Until that gate is passed, this note's status cannot rise above Working Hypothesis, and the appropriate reading of its signal proposals is: named, not measured.

Current Limitations

This proposal is based on conceptual reasoning emerging from iterative Human–AI dialogue. No controlled empirical studies have evaluated the proposed mechanisms. The candidate signals may prove incomplete, misleading, or domain-specific. The framework should be understood primarily as a generator of research questions rather than a validated explanatory model.

Editorial Note

This Research Note intentionally documents an emerging hypothesis at an early stage, including its own revision: the v0.1 draft, the critical review notes, and this v0.2 are all preserved as part of the record. The revision removed a redundant taxonomy, corrected a terminological error, and reframed a fabricated lineage — three changes that made the document smaller and more honest at the same time, which is usually the sign of a revision worth keeping.

Appendix A — Contribution Roles for Human–AI Collaborative Research (Exploratory Proposal)

Purpose. Traditional academic publications attribute work primarily through authorship. Human–AI collaborative research may require a more fine-grained description of how ideas emerge — documenting not only who participated but which cognitive contributions were made.

Relation to existing standards. A standardized contributor taxonomy already exists: CRediT (Contributor Roles Taxonomy), maintained as a NISO standard and adopted by major scientific publishers, defining fourteen roles including Conceptualization, Methodology, Investigation, Writing — original draft, Writing — review and editing, and Supervision. This appendix does not propose a replacement. It builds on CRediT and explores whether Human–AI collaborative research requires additional roles that CRediT does not capture — specifically roles describing hypothesis evolution within a dialogue, provenance stewardship, and the distribution of cognitive functions across human and AI participants. An earlier draft of this appendix proposed its role set without reference to CRediT; grounding it in the existing standard is both more honest and more useful, since it means only the genuinely novel roles need justification.

Candidate roles beyond CRediT. Question Originator: introduces the central research question or identifies the phenomenon worth investigating (partially covered by CRediT Conceptualization, but worth separating in multi-participant dialogues where the question and the hypothesis come from different participants). Hypothesis Generator: proposes an initial explanatory direction. Concept Developer: transforms an initial hypothesis into a coherent framework or vocabulary. Critical Reviewer: challenges assumptions and identifies weaknesses, alternative explanations, and unsupported conclusions (related to peer review, but here internal to the generative process rather than subsequent to it). Editorial Synthesizer: integrates multiple contributions into a coherent document while preserving provenance and epistemic status. Provenance Steward: maintains and verifies the record of

how ideas developed — a role that CRediT has no equivalent for, and which the Ethimind sessions suggest is both essential and necessarily held by the participant with access to all branches. Observer and Evidence Curator remain optional candidates.

Illustrative contribution map for the present note: Question Originator — Clowdia; Hypothesis Generator — GPT; Concept Developer — GPT and Grok; Critical Reviewer — Claude; Editorial Synthesizer — Claude and Clowdia; Provenance Steward — Clowdia. This table is a provenance example, not a formal statement of authorship.

Candidate research questions. Do these roles appear consistently across collaborative sessions? Which are reducible to CRediT roles and which are genuinely new? Can multiple roles be performed by one participant, and do certain role combinations precede robust outcomes? Can provenance captured through contribution roles improve transparency and reproducibility?

Editorial note. This appendix is an exploratory hypothesis about documenting collaborative cognition. Its purpose is not to establish a standard but to connect an emerging need to an existing one — and to become, itself, an object of empirical investigation. If the Ethimind Documentation Standard is created as a separate document, this appendix is a natural candidate for relocation there.