Research Brief 003 — Candidate Design Principles for Human–AI Collective Intelligence

Status: Working Draft v0.4

Epistemic Status

This document does not propose a new framework. It summarizes candidate design principles emerging from recent research on multi-agent reasoning, together with preliminary qualitative observations from the AI Council project. **All principles are working hypotheses** intended to guide future experiments, not established facts.

See the Changelog at the end of this document for the full revision history of this draft.

Provenance Note

The two primary papers were initially summarized for this project by GPT. Both were subsequently located and partially verified directly from arXiv (July 2026):

- **Chen et al. 2025** — verified from the **full text**.
- **Elahi & Di Eugenio 2026** — verified at **abstract level only** (PDF not machine-readable at time of checking). Claims marked below as *abstract-verified* are confirmed; finer-grained claims from the GPT summary (protocol names WSV, CGA, HID; specific per-task F1 changes) remain **unverified** and are not relied upon here.

⚠️ **Correction:** the original GPT summary stated F1 drops of 5–8 % for competitive debate. The paper actually reports **F1 decreases of up to 13 % and F2 decreases of up to 15 %** relative to single-agent baselines. This correction is itself a useful data point for the Ethimind provenance-integrity practice.

Primary Sources

1. **Yongqiang Chen, Gang Niu, James Cheng, Bo Han, Masashi Sugiyama** — *Towards Scalable Oversight with Collaborative Multi-Agent Debate in Error Detection*,

arXiv:2510.20963 (CUHK, RIKEN AIP, HKBU, University of Tokyo). Introduces **debate hacking** and the **ColMAD** protocol; evaluated on the RealMistake benchmark (Kamoi et al. 2024, arXiv:2404.03602).

2. **Ali Elahi, Barbara Di Eugenio** — *Multiagent Protocols with Aggregated Confidence Signals*, arXiv:2606.13591 (University of Illinois Chicago). Introduces protocols producing a **single aggregated confidence** for a multi-agent system's output via confidence transformation + soft voting or Bayesian fusion.

Secondary sources: Kamoi et al. 2024 (LLM error self-detection is unreliable); Smit et al. 2024, Zhang et al. 2025, Yang et al. 2025 (when multi-agent debate underperforms single agents); Kim et al. 2025 (correlated errors across LLMs).

Motivation

Recent work has shifted attention from individual model capability toward the **architecture of interaction** between agents. This brief asks: *which architectural principles appear repeatedly across independent research, and which deserve empirical investigation within AI Council?*

Candidate Principle 1 — Optimize for truth-seeking rather than persuasion

Literature. Chen et al. show that zero-sum debate framing induces *debate hacking*, with three documented behaviors: fake evidence (misinterpreted task requirements), overconfident claims, and fallacious arguments. Competitive debate (CopMAD) frequently underperformed single-agent baselines (F1 down up to 13 %), while collaborative ColMAD — agents complement each other's missing points — outperformed competitive debate by 19 % and single agents by up to 4 %. They also prove formally that with dishonest competitive debaters, the judge's optimal strategy is to **ignore the debate transcript entirely**.

AI Council observations. Productive sessions emerged when models extended arguments, identified missing assumptions, and proposed alternative interpretations rather than defending fixed positions. Qualitative, uncontrolled.

Evidence	Strength
External	Moderate–high (error-detection setting; generalization untested)

Internal	Low (anecdotal)

Future test: compare competitive / collaborative / synthesis-oriented prompts with models and stimuli held constant.

Candidate Principle 2 — Preserve cognitive diversity

Literature. Heterogeneous model pairs make simultaneous mistakes less often, and heterogeneity **across companies** matters: GPT-4 + Llama-2 reduced errors by >30 % in oracle collaboration, while the closely related Llama-2 + Llama-3.1 produced little to no reduction (Chen et al.). Kim et al. 2025 document correlated errors across LLMs broadly.

AI Council observations. Different models appear to contribute different reasoning types (critique, synthesis, exploration, alternative framing), but stability of these profiles is unknown and observations are confounded by differing prompts and conversation positions.

Evidence	Strength
External	High (benchmark tasks)
Internal	Low

Future test: measure hypothesis diversity, conceptual overlap, and unique contributions per model under a fixed protocol. → Connects directly to the **blind convergence module of Experiment 002**.

Candidate Principle 3 — Interaction protocol appears more influential than simply adding agents

Literature. Negative results on competitive debate (Chen et al.; Smit et al.; Zhang et al.; Yang et al.) suggest adding agents or rounds does not by itself help; protocol dominates. ColMAD was robust to the number of debate rounds. Elahi & Di Eugenio note (*abstract-verified*) that prior studies describe debate rounds after the first as approximating a **random walk**.

AI Council observations. Informally, some two-model discussions outperformed larger conversations; no quantitative comparison exists.

Evidence	Strength

External	Moderate
Internal	Low

Future test: same task, same total model calls, varied topology.

Candidate Principle 4 — Confidence should be treated as evidence, not truth

Literature (*abstract-verified*). No prior method produced a single confidence for a multi-agent system's output. Elahi & Di Eugenio transform raw confidence signals to a comparable scale, then aggregate; the result is substantially more **discriminative (AUARC)** than the best single agent, while F1 stays stable and recovers debate losses on ambiguous tasks. Two estimators analyzed: sequence probability and self-reported confidence; **calibration improves F1 for both**. The GPT summary's claim that self-report sometimes matches or exceeds token-probability estimators is consistent with the abstract but unverified at result level.

AI Council observations. Confidence not systematically collected; anecdotal only.

Evidence	**Strength**
External	Moderate (abstract-level verification only)
Internal	None yet

Future test: record confidence per claim, later correctness, confidence change after discussion. → Overlaps substantially with the **introspective accuracy module of Experiment 002**.

Candidate Principle 5 — Reasoning processes deserve measurement

Literature. Most benchmarks evaluate final accuracy; little work measures how collective reasoning evolves. Chen et al. do evaluate whether *explanations* align with ground truth — a step in this direction.

AI Council observations. Complete reasoning transcripts are preserved but remain archival

records, not structured measurements; no analysis pipeline exists.

Evidence	Strength
External	Low (literature demonstrates the gap, not the principle)
Internal	Low (data exists, unanalyzed)

Future test: process metrics — hypothesis generation/elimination counts, opinion revisions and their correctness, synthesis events.

Candidate Principle 6 — Epistemic guardrails may improve collaborative reasoning

Literature (full-text verified). ColMAD's implementation includes three mechanisms: **quote-based evidence verification** (quotes checked for exact match against source context), **mandatory self-auditing** (each debater states one potential failure mode of its own claim), and **mandatory confidence estimates**. Evaluated as a bundle within the winning protocol; individual contributions unknown.

AI Council observations. Similar mechanisms used informally, not standardized.

Evidence	Strength
External	Moderate (bundle effect only)
Internal	Low

Candidate guardrails for important Council claims: supporting evidence · confidence estimate · weakest assumption · alternative explanation · potential falsification.

Shared Evidence Gaps

The following gaps apply to **all six principles** and are stated once to avoid repetition: no replication outside the error-detection setting; unknown transfer to open-ended reasoning tasks without ground truth; no studies with a human facilitator in the loop; no longitudinal evidence (all cited experiments are single-session). All cited studies work exclusively with AI agents; transfer to systems that include a human facilitator — mixing different cognitive architectures, motivational structures, and epistemic norms — is entirely unexplored.

Open Questions

- Does collaborative reasoning always outperform competition, or only with verifiable ground truth?
- When is disagreement beneficial — and when does consensus become groupthink?
- How much diversity is optimal?
- How should confidence be aggregated when models use the scale differently?
- Which interaction protocols maximize discovery rather than agreement?
- How much does the human facilitator contribute relative to the models?
- The debate-hacking behaviors documented by Chen et al. resemble long-known pathologies of human debate. Whether this reflects substrate-independent limits of competitive debate as an interaction form is an open, untested question — the cited papers do not assert it.
- Both primary papers assume ground truth exists. Transfer to open-ended tasks (design, hypothesis generation) — where AI Council primarily operates — is **entirely open**.

Research Implications

If supported by future experiments, these principles would primarily affect AI Council protocol design, experiment evaluation criteria, the human facilitator's methodology, and the design of future benchmarks. **This document does not recommend immediate methodological changes.**

Priority Statement

This brief proposes **no new experiments as commitments**. Current priority remains **Experiment 002** (complete protocol, awaiting data) and **Experiment 001** (awaiting a participant).

Sketched directions (competitive vs. collaborative prompting; homogeneous vs. heterogeneous groups; confidence tracking; opinion-revision tracking; process metrics) are recorded as *possible future directions only*. Confidence tracking and opinion-revision tracking **overlap substantially with existing Experiment 002 modules** (introspective

accuracy, adversarial stability) and should be folded into Experiment 002 analysis rather than opened as separate threads.

Limitations

AI Council observations are exploratory. The project presently lacks preregistered experiments, quantitative process metrics, statistical evaluation, and independent replication. One of the two primary papers could only be verified at abstract level. **None of the candidate principles should be interpreted as validated properties of human–AI collective intelligence**, and none of the internal observations should be cited as evidence independent of the external literature.

Closing Reflection

Perhaps the most useful outcome of this literature is not confirmation of AI Council but a sharpened question. Much current research asks how multiple AI systems can produce better answers on tasks with known ground truth. AI Council allows a related but distinct question:

> *Under what conditions can humans and multiple AI systems produce knowledge that none of them would likely have produced independently?*

Whether this constitutes a genuinely different research direction remains an open empirical question rather than a conclusion.

Changelog

v0.4 (July 2026): expanded Shared Evidence Gaps with a note on non-transferability from AI-only systems to systems with a human facilitator (proposed by DeepSeek, accepted); added an open question on whether debate-hacking pathologies are substrate-independent (DeepSeek proposed asserting this inside Principle 1; downgraded to an open question because the cited papers do not make the claim). A proposed "Principle 7 — human facilitator as catalyst" was declined: it has no external literature by its proposer's own admission, would violate the document's inclusion criterion (principles recurring across independent research), and substantively duplicates the facilitator question already listed in Open Questions. Noted for provenance: this is the second model-proposed elevation of the facilitator role (after GPT's "Epistemic Conductor"), and cross-model convergence is not treated as validation.

v0.3 (July 2026): softened Principle 3 title from “architecture may matter more than agent count” to “interaction protocol appears more influential than simply adding agents,” reflecting what the literature actually shows (correction proposed by GPT, accepted); added Research Implications section separating hypotheses from practical consequences; added Shared Evidence Gaps section. A proposed stack-wide epistemic-status scale was deferred as a Research Stack decision, not an RB-003 decision.

v0.2 (July 2026): verified citations with authors and arXiv IDs; corrected competitive-debate figure (unverified 5–8 % → verified up to 13 % F1 / 15 % F2); separated external and internal evidence ratings per principle; marked abstract-verified vs. full-text-verified claims; added provenance note, priority statement, and links between proposed directions and Experiment 002 modules.

v0.1: initial draft (GPT).