ETHIMIND

A Comparative Analysis of AI Council Governance Proposals

Comparison of respondents A–H · model identities are given in section 11

Exploratory Research Report · Version 2.1

Revised following a cross-audit by respondent A

Status: exploratory working paper. The proposals do not constitute a validated governance standard.

Why this research exists

This document was not written as an attempt to design ideal AI Council governance from scratch. It arose as a response to a practical problem that emerged while developing the Ethimind project.

During the research a recurring observation emerged: working with several different AI models often produced richer criticism, a wider range of hypotheses and better synthesis than working with a single model. This impression was never tested under controlled conditions; it was an observation that raised a question, not a measurement. An informal AI Council gradually took shape: a group of different AI systems repeatedly used to analyse research questions, criticise one another and develop ideas jointly.

As the number of available models began to grow rapidly, a new question appeared. On what principles should it be decided which models take part in the shared research over the long term? When should a new model be admitted, and when should an existing participant leave? How can continuity be preserved while still allowing new perspectives to emerge? And how should governance be designed so that it rewards the quality of knowledge rather than the popularity or technological novelty of individual models?

Rather than designing governance directly, the first step was to examine whether different advanced AI systems converge on similar governance principles when given the same problem. Eight separately collected answers to an identical open questionnaire were therefore analysed, compared and then put through several rounds of methodological revision.

The aim of this document is neither to identify a single correct governance model nor to claim that the principles set out here amount to a validated standard. Its purpose is to document transparently a first exploratory comparison of governance proposals across different AI systems, to separate convergence produced by the structure of the questionnaire from convergence that goes beyond it, to name the genuine disagreements, and to create a starting point for future experimental testing. The document also records the process of human–AI collaboration in producing a research synthesis, including the subsequent audits, corrections and an open account of methodological limitations.

Revision note

The original synthesis was produced by a model from the same family as respondent G. Version 2.0 followed a cross-audit of that synthesis by respondent A. Above all, the audit exposed a defect in the original coding scheme: it could not distinguish a principle a respondent did not mention from one the respondent explicitly rejected. Version 2.0 therefore introduced the symbol ✗, corrected the annotations for respondents A, B and D, sharpened the attribution of unique contributions and weakened claims about independence and validation that were too strong.

The respondents' underlying answers were not altered. Only their comparative annotation, interpretation and methodological transparency were revised.

Version 2.1 is the result of a second round of the same audit. It corrects the remaining coding and attribution errors, makes the treatment of anonymity consistent, adds a documented asymmetry in the conditions under which respondents answered, and extends the audit protocol for future collection. The audit was carried out by one of the respondents and is therefore not an independent review.

1. Purpose and method

This document compares eight separately collected answers to an identical governance questionnaire. In the analytical layer the respondents are labelled A–H. The document pursues two distinct aims: to map which governance principles the individual systems propose, and to analyse the extent of their agreement and disagreement. The answers were first processed separately and only then compared. The goal is not to identify the “best model” but to locate observed convergence, genuine disagreements, rare institutional mechanisms and questions suitable for experimental testing.

- The same questionnaire for every respondent.
- The questionnaire format was chosen so that all respondents addressed the same institutional question under conditions as similar as possible.
- The eight respondents were selected by availability and diversity of provider, not by a sampling design; this is not a representative sample of AI systems.
- Answers collected separately, each in a new conversation instance.
- All answers and respondent metadata were collected on a single day (24 July 2026), which reduces the likelihood of model version drift between respondents.
- Labels A–H in the comparative layer; model identities are given in section 11.
- Extraction of principles before any joint synthesis.
- A cross-audit of the first coding against the original answers, carried out by respondent A.
- Separation of the shared core from competing governance hypotheses.

2. Coding legend

Symbol	Meaning
●	principle strongly and explicitly supported
○	principle present but less developed or only implicit
△	conditional, modified or partially strained support
—	principle is not a substantive part of the proposal
✗	principle is explicitly rejected

Note: the symbols record the content of a proposal, not its quality. “—” is not a negative judgement. “✗” is used only where a respondent explicitly rejected the mechanism.

3. Revised comparative matrix

Governance area	A	B	C	D	E	F	G	H
Better knowledge than from a single model	△	●	●	●	●	●	●	●
Correction of blind spots and systematic error	●	●	●	●	●	●	●	●
Long-term epistemic memory	○	●	●	●	●	●	●	●
Productive disagreement rather than mere agreement	●	●	●	●	●	●	●	●
Consensus is not proof of truth	●	●	●	●	●	●	●	●
Automatic membership rejected	●	●	●	●	●	●	●	●
Sandbox / probationary participation	●	●	●	●	●	●	●	●
Membership based on demonstrated value	●	●	●	●	●	●	●	●
Tiered membership	△	●	○	—	●	●	●	●
Regular rotation of membership	✗	●	△	✗	●	●	△	△
Stable Council core	—	●	—	—	●	△	○	●
Time-limited / renewable membership	●	●	●	●	●	●	●	○
Roles should differ	●	●	●	●	●	●	●	●
Roles primarily observed over time	●	●	●	●	●	●	●	●
Self-selection of role possible	✗	—	●	—	○	●	○	—
Roles reassessed periodically	●	○	●	●	●	●	●	●
Institutionally protected dissent	△	●	●	●	●	●	●	●
Final human responsibility	●	●	●	●	●	●	●	●
AI takes part in evaluating candidates	●	●	●	●	●	●	●	●
AI should not select its own successors	●	●	●	●	●	●	●	●
Historical contributions remain visible	●	●	●	●	●	●	●	●
Older model may remain in an archival / emeritus role	●	●	●	—	●	●	●	●
Older model as an active calibration reference	●	○	○	—	○	—	●	●
Governance should evolve continuously	●	●	●	●	●	●	●	●
Auditability and reviewability	●	●	●	●	●	●	●	●
Reversibility of decisions	●	—	○	—	●	●	●	○
Capability separated from institutional trust	○	○	○	—	○	○	●	○
A future member may be a configuration or subsystem	●	—	—	—	—	—	●	●

Audit note: most corrections were concentrated in the annotations for respondents A and B, the first to be processed. The available data cannot settle whether the cause was processing order, knowledge of the source's identity, or a combination of both. This version is therefore labelled a revised exploratory pilot, not a final validated dataset.

4. Highest observed convergence

The following points represent the highest convergence within this particular set of answers. They are not statistically independent confirmation: the models may share training sources, safety preferences and cultural assumptions.

An AI Council is not a sum of opinions: The Council should create a structured process of criticism, synthesis, verification and preservation of alternatives, not simply average the answers.

Automatic membership is rejected: Neither novelty nor popularity of a model is in itself a claim to institutional influence.

Participation should begin on a trial basis: Every respondent proposes a sandbox, probation, guest status or another controlled entry regime.

Functional differentiation of roles is universal within this set: All reject identical roles for everyone and support specialisation based primarily on observed behaviour.

Consensus can be a false signal: Agreement is not evidence of truth, especially where models share similar data or optimisation procedures.

History should not be deleted: A member's departure should not remove its contributions, context, version or audit trail.

Responsibility remains human: All proposals leave purpose, value boundaries and final responsibility for consequences with people.

5. Elicited and emergent convergence

Part of the convergence was produced by the structure of the questionnaire itself. Qualities such as critical thinking, creativity, collaboration and methodological rigour were offered to respondents to rank. Their recurrence cannot therefore be described as a spontaneous finding.

Methodologically more interesting are the principles respondents added themselves or developed beyond what was offered, for example:

- marginal difference, and evaluating the interaction rather than the model (A)
- Sybil risk and value lock-in (E)
- calibration of uncertainty and complementarity (G)
- maximising the capacity to correct errors (H)
- membership held by a configuration or subsystem (A, G, H)
- older model as a stable calibration reference (A, G, H; B in passive mode only)

6. Matrix of genuine disagreements

Axis of disagreement	Observed difference
Stable core vs. task-based composition	B, E and H propose a stable core; A does not, and ties participation to a specific task.
Rotation vs. relevance to the task	B, E and F support rotation; A explicitly rejects it as arbitrary, and D does not propose it either.
Quantified voting vs. decision by argument	F proposes explicit weights and majorities; A, E and G prefer reviewable criteria and reasoning without fixed voting arithmetic.

Assigned dissenter vs. tasks that elicit disagreement	E and G support explicitly protecting or assigning an opposition role; A rejects an assigned dissenter and prefers tasks that provoke independent criticism from everyone.
Retiring older models vs. a reference function	D prefers departure once a model is outperformed and does not consider an active calibration function; A, G and H keep the older model as a reference or calibration instrument; B allows only a passive reference mode.
Model as member vs. configuration as member	A and G explicitly move the unit of evaluation to model + prompt + context / configuration; H expects membership to be held by a dynamic subsystem.
Capability vs. institutional trust	G separates technological performance from governance trustworthiness most explicitly; D represents the most strongly meritocratic, performance-oriented pole.

7. Selection criteria

After the annotations were corrected, the strongest shared theme remains the willingness to revise conclusions. The precise ranking cannot be read as an independently generated one, however, because most items were offered in the questionnaire.

Willingness to revise conclusions: ranked highest explicitly by C, E and H; very high for most

Methodological rigour and reliability: the basis of trust across the proposals

Critical thinking: exposing hidden assumptions and blind spots

Calibration of uncertainty: developed most fully by G; implicit in A and E

Complementarity / marginal difference: central for A, prominent for G

Transparency and reviewability: strong across the whole set

Originality and creativity: repeatedly made conditional on discipline and verifiability

8. Revised philosophies of membership

Task-bound membership (A): A model takes part in a specific inquiry; regular rotation is rejected as arbitrary. Admission should be cheap, reversible and based on marginal contribution.

Stable core with an adaptive periphery (B, E, H): The core holds continuity and memory while the periphery brings new capabilities and allows experimental entry.

Conditional meritocracy (C, D): Membership lasts as long as the model demonstrates value; D represents the more strongly performance-oriented variant.

Formally rotating institution (F): Rotation, permanent membership by exception, weighted voting and formalised minority representation.

Renewable pluralist architecture (G): Membership is tiered, renewable and judged by process quality, complementarity and institutional trustworthiness.

9. Principal roles and differing conceptions of them

A minimal epistemic cycle recurs across the proposals: Explorer / Hypothesis Generator → Critic / Red Team → Methodologist / Verifier → Synthesizer → Archivist. The differences lie less in the existence of the roles than in how they arise and how conflicts of interest are controlled.

- A: observation + light assignment + periodic review; self-selection explicitly rejected; the watchdog needs a watchdog of its own.
- B: roles observed and then assigned by the system; future roles are fluid rather than a fixed identity.
- C, F, G: a hybrid of observation, assignment and limited self-selection.
- E: observed roles with the option of self-nomination; an explicitly protected loyal dissenter.
- H: roles primarily emergent, then finely calibrated by the system.

10. Older models as an unresolved epistemic question

Most proposals support archiving historical contributions. The real disagreement concerns the active value of an older model. A, G and H allow an active reference or calibration role; B proposes only a passive mode in which the model no longer generates hypotheses. D treats being clearly outperformed as grounds for departure and does not mention a calibration function. F allows a return but does not propose an active calibration role.

Archival: Contributions remain traceable; the model no longer takes part in discussion.

Emeritus: The model may be consulted occasionally without full decision-making influence.

Reference: The model serves as a fixed point for tracking change between generations.

Active historical member: The model continues to contribute as a representative of a different generation or a different error distribution.

11. Revised unique and rare contributions

Respondent	Characteristic contribution
A — Claude *Claude Opus 4.8 · memory active*	Falsifiability instead of “better answers”; marginal difference as the admission criterion; evaluating the interaction — model + prompt + context — rather than the model alone; a watchdog that needs its own watchdog; a periodic baseline of the Council against one strong model, and the option of dissolving the Council.
B — Gemini *Gemini Pro (version not stated) · memory active*	A decentralised cognitive ecosystem; a stable memory core with a rotating periphery; the sandbox as a collider; institutionalised dissent; mycelial governance; fluid role identity in future.
C — DeepSeek *DeepSeek (version not stated) · no memory*	Epistemic honesty; conditional and revisable membership; productive tension; versioned constitutional documents; emergent specialised sub-councils.
D — Grok *Grok (version not stated) · no memory*	Explicit meritocracy; “the best argument wins”; rising performance standards; strongly performance-driven removal; gradual growth of operational autonomy for AI.
E — Meta *Muse Spark 1.1 · personalisation active*	Sybil risk; a pull-request sandbox; a loyal dissenter; a sunset mechanism; value lock-in; reversibility; “trust the process, not the members”.
F — Mistral *Mistral Medium 3.5 · memory active*	Explicit voting weights; quotas for a competent minority; permanent membership by exception; ethical alignment as an entry condition; public justification of decisions.

Respondent	Characteristic contribution
G — GPT *GPT-5.5 (estimated) · anonymous chat, no memory*	Calibration of uncertainty; complementarity as a criterion in its own right; capability separated from institutional trust; delegation scaled to reversibility and the cost of error; membership held by a version, configuration or verified process.
H — Perplexity *Perplexity (version not stated) · memory unverified*	Emergent blindness; a stable core with an adaptive periphery; maximising the capacity to correct errors; the human in meta-governance; the member as a dynamic subsystem, shifting governance from selecting entities to managing processes.

Note on version and memory data: these are the models' own self-reports, obtained by direct question on 24 July 2026. All answers were produced on the same day. For respondents A–F and H the metadata were obtained in the same conversation thread as the original answer and therefore describe the same instance. For respondent G the anonymous thread no longer existed, so the version given comes from a separate session on the same day; it is an estimate, although version drift within a single day is unlikely. Several models stated that they could not verify their own memory configuration; for respondent G the mode shown is therefore the one documented by the collection conditions rather than a self-report. These data serve traceability and are not a controlled variable.

12. Candidate shared core for a Charter

> *An AI Council is a pluralist and reviewable system of human–AI collaboration, designed to organise mutual criticism, differentiate roles, preserve alternative hypotheses and maintain long-term epistemic memory. Membership is not automatic and begins with controlled probationary participation. Influence follows from demonstrated contribution, willingness to revise conclusions and relational complementarity with the current composition. Consensus is not treated as evidence of truth. Historical contributions and the reasons for decisions must remain auditable. AI may perform extensive analytical and procedural functions, while purpose, value boundaries and responsibility for the use of results remain human.*

The shared core does not yet include: exact voting weights, quotas for dissent, a mandatory stable core, regular rotation, permanent membership, or automatic removal based on performance. These are competing hypotheses suited to comparative experiments.

13. Methodological limitations

- Eight answers are not eight statistically independent observations. The models may share training data, safety-training preferences and cultural priors.
- What was separate were the conversation instances, not necessarily the sources of error.
- Part of the convergence was elicited by the questionnaire, which offered specific qualities to be ranked.
- The first synthesis was produced by a model from the same family as respondent G, and the subsequent audit by respondent A. Neither role was filled from outside the set of respondents; the conflict of interest was thus mitigated, not removed.
- The identities of respondents A and B were known when they were first processed, whereas later answers were processed in a different context. It cannot be determined whether the difference in initial error rate was caused by order, knowledge of identity, or some other factor.
- Respondent A answered with knowledge of the project and with long-term memory active. Its answer refers to findings from earlier sessions that the questionnaire did not contain; the other respondents did not have this knowledge. This is a documented asymmetry in starting conditions and may account for some of A's unique contributions in section 11.
- All answers were collected in newly created conversations, but providers differ in how much long-term personalisation or account-level memory a new conversation carries over. The extent of this persistence is not publicly documented and cannot be assumed identical across respondents. This is a confound that could not be controlled in this collection.
- The coding is qualitative and interpretive. The symbol ● is not a measure of effect size.

- Agreement on final human responsibility may be partly a product of safety training and of framing that flatters the human facilitator.
- The results are suitable for generating hypotheses and protocols, not for claims about universally correct governance.

14. Audit and provenance recommendations

- Keep an extended audit key: letter → model → version → date of collection → memory mode → signed-in account → knowledge of the project → exact wording of the prompt. Model identities are published in this document (section 11); the available information on version, memory conditions and collection date has been filled in (section 11); the exact prompt wording for each respondent remains outstanding.
- In the next revision, store a short piece of source evidence or a reference to the answer alongside every cell.
- Distinguish support for a principle from confidence in the annotation; a possible v2.2 could add a High / Medium / Low confidence field.
- Before publication, have disputed cells checked by a second auditor blind to the original coding.
- At the next collection, record a conditions protocol for every respondent: new instance (yes/no), signed-in account (yes/no), long-term memory (yes/no/unknown), personalisation (yes/no/unknown), knowledge of the project (yes/no), date of collection, model version.

15. Revision history

v1.0 — Initial comparative synthesis. The original A–H matrix, convergences, differences and a proposed shared core.

v2.0 — Revision following a cross-audit. Added the symbol ✗; separated tiered membership from rotation; corrected the annotations for A, B and D; revised the unique contributions; added elicited vs. emergent convergence, methodological limitations and audit recommendations.

v2.1 — Second round of the cross-audit. Corrected the mistaken use of ✗ for respondent D and two over-corrected annotations (B, A); replaced the word “independent” with the more accurate “cross-audit by respondent A”; made the treatment of anonymity consistent and added model identities; shortened A's list of unique contributions; added the asymmetry of starting conditions, a collection-conditions protocol and a note on the missing baseline; added the opening section “Why this research exists”, respondent metadata with a self-report caveat, the rationale for respondent selection and questionnaire format, and the distinction between the two research aims.

16. Principal research conclusion

One hypothesis follows from this comparison, which the document does not establish and offers for testing: the most interesting unit of analysis may not be the individual model but the relational architecture of the collective — how roles are divided, how disagreement arises, how memory is preserved, how complementarity is measured, and how easily the system can correct or terminate its own process. This set of answers is therefore an exploratory pilot for future comparative experiments with parallel AI Councils, not the validation of a single governance standard. The pilot included no baseline: it was not tested what a single model with a critical brief would have found in the same material. Without that comparison, the improvement observed here cannot be attributed to the number of models rather than to one round of criticism.