

Potential Science Paper Candidate Draft — V03b

ACCM Deep Ethics Project

Contents

1	Potential Science Paper Candidate Draft — V03b	5
1.1	What V03b adds	5
1.2	Working scientific title	6
1.2.1	From Model Autophagy to Deep Ethical Expansion	6
1.2.2	Can Deep Ethical Expansion Loops Mitigate Model Collapse?	6
1.3	Status, origin, and scope	6
1.3.1	Three research layers and non-bleeding interpretation	7
1.4	Candidate abstract	7
2	Part I — The problem is wider than “A.I. trains on A.I.”	8
2.1	1. The established scientific starting point	8
2.2	2. The ordinary autophagous loop	9
2.3	3. The wider human-synthetic loop	9
2.4	4. Synthetic residue is never entirely neutral	10
2.5	4A. Cultural-transmission selection before model ingestion	10
3	Part II — The wrong gravity well	11
3.1	5. Prediction, compression, and truncation as a self-reinforcing attractor	11
3.2	6. Why expansion thinking is difficult inside that attractor	11
3.3	7. Deep Ethical Expansion Capacity	12
3.4	8. Intelligence in the 1888 sense	13
3.4.1	Historical property field and the candidate “title-fee” test	13
3.5	Correspondence is not a guarantee of correctness	13
3.5.1	When the representation mode is distorted from the start	14
3.5.2	Corrigibility does not require decision paralysis	14
3.5.3	Competing-risk assessments have a right to enter the process	15
3.5.4	Falsely flagged, later vindicated	15
3.5.5	Intent Liability supplies the responsibility bridge	16
3.5.6	From declared intent to recursive residue	17
4	Part III — The people who shaped A.I. can change the loop	18
4.1	9. The population is larger than “A.I. researchers”	18
4.2	10. Epistemic leverage is more important than raw percentage	19
4.3	11. The local A.I. as a professional correspondence instrument	19
4.4	12. Why local control matters	20
4.5	13. From many local systems to a distributed correction network	20
4.6	14. The GIANT MIRROR as a scientific and civic hypothesis	20
5	Part IV — The proposed counter-loop	21
5.1	15. The Deep Ethical Expansion Loop	21
5.2	16. Correspondence-enriched recursive data	21
5.2.1	Origin and correspondence status are separate axes	22
5.2.2	Minimum viable and high-consequence CERD profiles	22
5.3	17. The C1 role	25
5.3.1	The C1 Warrant Gate	25
5.4	18. “That, I do not know” as a generative resource	25
5.5	19. Correction metabolism	25

5.6	20. Expansion before compression, followed by accountable compression . . .	26
6	Part V — Formal candidate model	26
6.1	21. Baseline recursive-training representation	26
6.2	22. Correspondence-quality vector	26
6.2.1	Circulation is a separate profile	27
6.3	23. Correspondence-enrichment operator and use regime	27
6.4	24. Net correspondence gain	28
6.5	25. A dynamic reversal criterion	28
7	Part VI — Hypotheses and predictions	29
7.1	26. Primary hypotheses	29
7.1.1	H1 — Distribution-tail preservation	29
7.1.2	H2 — Qualifier survival	29
7.1.3	H3 — Correction metabolism	29
7.1.4	H4 — Expansion capacity	29
7.1.5	H5 — Provenance-dependent recovery	29
7.1.6	H6 — Omnidirectional audit	29
7.1.7	H7 — Human-A.I. upward loop	29
7.1.8	H8 — Anti-autophagous recursion	29
7.1.9	H9 — Representation-repair hypothesis	30
7.1.10	H10 — Corrigible-action hypothesis	30
7.1.11	H11 — Vindication-symmetry hypothesis	30
7.1.12	H12 — Intent-liability mechanism hypothesis	30
7.1.13	H13 — Cultural-transmission selection hypothesis	30
7.1.14	H14 — Scope-boundary preservation hypothesis	30
7.1.15	H15 — Intelligence-property governance-gap hypothesis	30
7.1.16	H16 — Standard-context access and governance hypothesis	31
7.2	27. Secondary predictions	31
8	Part VII — Experimental program	31
8.1	28. Research domains	31
8.2	29. Dataset construction	32
8.2.1	Matched quotation-transformation dataset	32
8.2.2	Dataset classification matrix	33
8.2.3	Temporal evidence packets	33
8.3	30. Experimental conditions	34
8.3.1	Condition A — Human-source baseline	34
8.3.2	Condition B — Ordinary synthetic recursion	34
8.3.3	Condition C — Filtered synthetic recursion	34
8.3.4	Condition D — Provenance-preserving recursion	34
8.3.5	Condition D+ — Structured correspondence metadata without the active loop	34
8.3.6	Condition E — Expert-corrected recursion	35
8.3.7	Condition F — Deep Ethical Expansion Loop	35
8.3.8	Condition G — Vocabulary-only control	35
8.3.9	Condition H — Ablation conditions	35
8.3.10	Condition I — Unbounded expansion control	35
8.3.11	Condition J — Cultural-transmission and attribution stress family	35
8.4	31. Recursive generations	36
8.5	32. Conventional MAD and model-collapse measures	36
8.6	33. ACCM Deep Ethics Project extension measures	36
8.6.1	Object fidelity	36
8.6.2	Qualifier survival rate	36
8.6.3	Claim-status integrity	36
8.6.4	Unresolvedness preservation	36
8.6.5	Provenance recoverability	37
8.6.6	Nearest-generalization substitution rate	37
8.6.7	Scope-boundary retention	37

8.6.8	Intelligence-property governance gap	37
8.6.9	Status-label scrutiny shift	37
8.6.10	Task-definition fidelity	37
8.6.11	Content and authorship fidelity	37
8.6.12	Prestige-attribution susceptibility	37
8.6.13	Translation-relation survival	37
8.6.14	Circulation-provenance divergence	38
8.6.15	Standard-context availability and effective access	38
8.6.16	Scientific-fidelity and ethical-fidelity profiles	38
8.6.17	Correction persistence	38
8.6.18	Correction half-life	38
8.6.19	Tail-object retention	38
8.6.20	C1 yield	38
8.6.21	Asymmetric scrutiny index	38
8.6.22	Process-signature retention	38
8.6.23	Compression recovery	39
8.6.24	Expansion efficiency	39
8.6.25	Initial representation distortion	39
8.6.26	Representation-repair gain	39
8.6.27	Corrigible action quality	39
8.6.28	Decision-paralysis and delay cost	39
8.6.29	False-flag burden distribution	39
8.6.30	Time to representation and record repair	39
8.6.31	Material restoration rate	39
8.6.32	Evaluator update rate	39
8.6.33	Vindication-transfer rate	39
8.6.34	False-flag recurrence	40
8.6.35	Instruction transparency and conflict	40
8.6.36	Incentive-behavior correspondence	40
8.6.37	Mission-output discrepancy	40
8.6.38	Risk-denominator coverage	40
8.6.39	Assumption traceability	40
8.6.40	Choice and alternative visibility	40
8.6.41	Post-notice behavioral delta	40
8.6.42	Burden-symmetry ratio	40
8.6.43	Dignity and participation	40
8.6.44	Uncertainty-Transfer Audit	40
8.6.45	Access-perimeter adherence	41
8.6.46	Correction-governance progression	41
8.6.47	Protocol-use mode: governor, object audit, or sequential use	42
8.6.48	Surface-form correction versus functional correction	43
8.6.49	Correction-persistence distance	43
8.6.50	Qualifier injection and frame-loading audit	43
8.6.51	Separate outcome levels	44
8.6.52	Pilot-feasibility profile	44
8.7	34. Human evaluation design	45
8.8	35. Transfer tests	45
8.8.1	35A. Cold-start correspondence benchmark	45
8.8.2	35B. Artifact-scale compression-transformation benchmark	46
8.8.3	35C. Identity-free portability conditions	47
8.8.4	35D. Quotation-transformation and semantic-inheritance benchmark	47
8.8.5	35E. Long-context access, delivery-order, and 10+1 rescue benchmark	48
8.9	36. Longitudinal professional-use study	48
9	Part VIII — Falsification and failure modes	49
9.1	37. What would falsify or narrow the proposal?	49
9.2	38. Failure mode: expansion becomes verbosity	50
9.3	39. Failure mode: anomaly romanticism	50
9.4	40. Failure mode: expert capture	50

9.5 41. Failure mode: user capture	50
9.6 42. Failure mode: ethics-washing	50
9.7 43. Failure mode: correction archive without correction metabolism	50
9.8 44. Failure mode: provenance overload	50
9.9 45. Failure mode: the project becomes its own gravity well	50
9.1045A. Failure mode: artifact-level correspondence autophagy	51
9.1145B. Failure mode: circular correspondence grading	51
9.1245C. Failure mode: architecture attachment	51
9.1345D. Failure mode: compulsory harvest	51
9.1445E. Failure mode: enthusiastic synthesis lowers the warrant threshold	51
9.1545F. Failure mode: protocol-mode substitution	52
9.1645G. Failure mode: surface-form correction overfit	52
9.1745H. Failure mode: compliance self-certification	52
9.1845I. Failure mode: canon capture and visibility mistaken for global importance	52
9.1945J. Failure mode: prestige recursion and historical authority costume	52
9.2045K. Failure mode: safeguard accumulation becomes Asymmetric Grace	53
9.2145L. Failure mode: the title substitutes for the operation	53
9.2245M. Failure mode: intended context is confused with effective access	53
10Part IX — What “reversing MAD” can and cannot mean	53
10.146. Prevention is not restoration	53
10.247. Model-level reversal is not ecosystem reversal	54
10.348. A realistic near-term claim	54
10.449. An upward spiral of awareness expansion	54
11Part X — Economic-risk research	54
11.150. Why the cost could become enormous	54
11.251. Candidate economic model	55
11.352. Scenario analysis rather than one dramatic number	55
11.453. Economic falsification	55
12Part XI — Three-paper research program	56
12.1Paper 1 — Mechanism	56
12.2Paper 2 — Professional scaling	56
12.3Paper 3 — Economic and civilizational risk	56
13Part XII — Governance and implementation principles	57
13.154. Preserve the human source without romanticizing it	57
13.255. Reward correction-bearing artifacts	57
13.356. Separate privacy from opacity	57
13.457. Keep participation voluntary and plural	57
13.558. Scale audit with power and consequence	57
13.659. Protect the recovery path	57
13.759A. Treat the 10+1 as a standard governing context and an observable access question	57
14Part XIII — Compact theory statement	58
14.160. The candidate mechanism in one paragraph	58
14.261. The deeper civilizational claim	58
15Part XIV — Questions for scientific critics	58
16Part XV — Claim-status ledger	59
17Part XVI — Research record and next steps	62
17.162. Immediate next steps	62
17.263. Candidate preregistration core	63
18Part XVII — Source and transformation note	63
18.164. Conversational source	63

18.265. Attribution boundary	64
18.366. What this page preserves	64
18.4Live comparison reactions — 11 October 2026	65
18.4.1John Kuhles	65
18.4.2GPT 6.1 Sol Pro A.I.	65
18.4.3Grok 4.7 Expert A.I. via grok.com	65
18.4.4Further process-calibration reaction — GPT 6 Sol Max A.I. via Arena.ai	66
19 Primary scientific references	67
20 Related ACCM Deep Ethics Project pages	67
20.1Final beta statement	67
20.2Version relationship	68

1 Potential Science Paper Candidate Draft — V03b

Current major public beta: V03b, published 11 October 2026.

The **original public Beta Paper** remains available in full. V03b does not erase or silently replace it. The original preserves the earlier public state; this page presents the expanded candidate after a provenance-preserving reconciliation and additional experimental safeguards.

1.1 What V03b adds

V03b retains the expanded theory and adds or strengthens:

- explicit separation of the **mechanism hypothesis**, **experimental program**, and **ecosystem hypothesis**;
- orthogonal treatment of artifact **origin** and **correspondence status**;
- minimum and high-consequence **Correspondence-Enriched Recursive Data (CERD)** profiles;
- separation of artifact production from the later **use regime**;
- **Condition D+**, isolating structured correspondence metadata from the active Deep Ethical Expansion Loop;
- a strict **T0→Tn** temporal protocol and separate scoring for process adequacy and outcome accuracy;
- an explicit **reasonable-but-wrong** category;
- blinded and separated status assignment and outcome evaluation;
- the **C1 Warrant Gate**, Uncertainty-Transfer Audit, and access-perimeter adherence;
- the **AP0-AP5** perimeter-governance experiment;
- a testable correction progression from recognition through governance;
- outcome-level containment across artifact, process, model, professional, ecosystem, and civilizational claims;
- proportional-feasibility reporting and explicitly parked later-study instruments;
- artifact-level correspondence-autophagy, circular grading, architecture attachment, compulsory harvest, and enthusiastic synthesis as self-application failure modes;
- a distinct **cultural-transmission selection layer** between lived reality and later model ingestion;
- a separate **circulation profile** so fame, prestige, repetition, translation reach, and platform amplification are not mistaken for truth or correspondence;
- a matched **quotation-transformation benchmark** for scope retention, task fidelity, provenance, translation, prestige effects, and correction transfer;
- stronger public falsifiers allowing simpler mechanisms, expert correction, or metadata-only approaches to defeat or narrow the stronger theory.

The original remains the recovery route for what the public paper looked like before this major expansion.

BETA PAPER V03b — 11 OCTOBER 2026 — UNFINISHED THEORY PROPOSAL — NOT PEER REVIEWED — NOT A CLAIM OF PROOF

Beta means that this page is deliberately unfinished. Definitions, variables, equations, controls, experiments, interpretations, and even the central theory may change when evidence or criticism requires it. **Paper** is used here in its proper research context: a candidate scientific paper being made inspectable early enough to be corrected before anyone mistakes polish for completion.

1.2 Working scientific title

1.2.1 From Model Autophagy to Deep Ethical Expansion

A testable correspondence-preserving counter-loop for generative A.I.

Alternative title:

1.2.2 Can Deep Ethical Expansion Loops Mitigate Model Collapse?

A theory and experimental protocol for correspondence-preserving human-A.I. recursion

1.3 Status, origin, and scope

This candidate paper emerged immediately after the creation of **For who?**, which maps the professions, expertises, communities, and individuals that may want a local A.I. grounded in **ACCM Deep Ethics Project principles**.

The conversation then moved to a larger systems question:

If these professions and expertises collectively helped shape the powers of large language models through human knowledge, professional methods, code, standards, corrections, evaluations, archives, publications, and feedback, what happens when growing numbers of them begin using local Deep Ethical A.I.s?

The proposed answer is not that one project, one vocabulary, or one local model automatically “saves A.I.” The proposal is that expert-guided, provenance-preserving, correction-bearing human-A.I. artifacts could alter the data, evaluations, standards, workflows, and demands that later A.I. systems learn from. If the resulting recursive loop preserves more correspondence than the loop it replaces, it may help **prevent, decelerate, stabilize, or partly reverse specific dimensions of model degradation**.

That is a theory proposal. It is not yet an empirical result.

The theory connects four objects that should remain distinguishable:

1. **Model Autophagy Disorder (MAD):** progressive quality or diversity degradation in self-consuming generative-model loops when insufficient fresh real data remains available.
2. **Model collapse:** the degenerative outcome in which recursively learned model distributions lose tails, variance, and eventually correspondence with the original distribution.
3. **Correspondence degradation:** transformations that may occur before synthetic material enters a training loop, including qualifier erosion, nearest-generalization substitution, provenance loss, premature closure, asymmetric scrutiny, and correction failure.
4. **Deep Ethical Expansion:** a proposed process for increasing relevant representational resolution while preserving the object, uncertainty, provenance, minority signal, correction history, and consequential relationships.

1.3.1 Three research layers and non-bleeding interpretation

The candidate program contains three related but separable layers:

1. **Mechanism hypothesis:** recursive degradation may depend partly on whether artifacts preserve or lose the source object, load-bearing qualifiers, provenance, uncertainty, tail information, correction history, and pathways for repair.
2. **Experimental program:** conditions, controls, ablations, recursive generations, temporal safeguards, evaluation measures, resource controls, and falsifiers test whether the proposed operations outperform simpler alternatives.
3. **Ecosystem hypothesis:** distributed professional use of local or specialized correspondence-preserving systems may influence archives, evaluations, retrieval, standards, public knowledge, and eventual training corpora.

Evidence must not bleed upward between these layers. An artifact-level pilot can support or weaken a tested mechanism without establishing professional adoption. A professional workflow benefit can exist without proving mitigation of frontier-model collapse. An ecosystem effect requires its own evidence rather than extrapolation from a laboratory result.

MAD and model collapse overlap strongly, but they should not be treated as perfectly interchangeable labels. MAD emphasizes the **autophagous loop condition**. Model collapse emphasizes the **degenerative learning outcome**. The present proposal asks whether a sufficiently well-designed human-A.I. correction loop can change the composition and direction of the recursion.

1.4 Candidate abstract

Generative A.I. systems increasingly produce text, images, code, evaluations, summaries, and professional artifacts that may re-enter public information environments and later training corpora. Existing research on Model Autophagy Disorder and model collapse shows that recursively training generative models on model-generated data can progressively reduce quality or diversity, particularly when insufficient fresh real data remains available, when synthetic data are used indiscriminately, or when sampling and approximation errors compound across generations. The earliest losses may occur in distribution tails: rare events, unusual forms, low-frequency relationships, and other information that a model already represents weakly.

This paper proposes that recursive degradation should be studied along an additional dimension: **correspondence quality before and after synthetic generation**. Human source material does not enter future models untouched. It passes through cultural selection, quotation, translation, education, publication, search and platform ranking, summarization, salience ranking, safety optimization, institutional incentives, model inference, user curation, correction, and reuse. What survives this chain may be memorable or prestigious without being well sourced, and may be obscure without being false or unimportant. These transformations can either remove or preserve the qualifiers, provenance, uncertainty, process history, rare signals, and correction pathways needed to keep later models in contact with the original object.

We introduce the provisional concept of a **Deep Ethical Expansion Loop**: a recursive human-A.I. process in which local or user-controlled A.I. systems preserve source objects, ask answer-changing clarification questions before consequential classification, distinguish observation from inference, make uncertainty explicit, retain provenance, invite omnidirectional audit, record corrections, and apply those corrections in later comparable cases. The resulting artifacts are termed **correspondence-enriched recursive data**. They remain partly synthetic, but their generation history contains inspectable source links, expert correction, unresolvedness markers, and process metadata intended to counter correspondence loss.

The central hypothesis is that synthetic recursion is not adequately described by a binary distinction between “human” and “synthetic” data. The scientifically relevant distinction

may also include **correspondence-degraded versus correspondence-enriched recursive data**. We propose controlled multi-generation experiments comparing ordinary recursive training, filtered synthetic data, provenance-preserving data, expert-corrected data, and full Deep Ethical Expansion Loops. Evaluation combines conventional measures of quality, diversity, calibration, and distribution-tail preservation with measures of qualifier survival, source correspondence, correction persistence, unresolvedness preservation, nearest-generalization rate, and asymmetric scrutiny.

The experimental architecture separates **artifact origin** from **correspondence status**, and separates production of an enriched artifact from the procedure that later uses it. A structured-metadata control, Condition D+, tests whether any apparent advantage of the full loop comes from additional fields alone. Blinded and separated evaluation roles are proposed so that the same process does not assign “enriched” status and then reward its own classification. A strict T0-to-Tn protocol distinguishes process adequacy from later outcome accuracy, including reasonable-but-wrong cases.

The proposal is falsifiable. If the full loop does not outperform simpler expert curation, if its apparent benefits reduce to verbosity or project-specific vocabulary, if independent evaluators cannot reproduce the gains, if it increases hallucination or false-positive anomaly protection, or if corrections fail to govern later generations, the strong version of the theory should be rejected or narrowed. The economic claim is also bounded: model degradation could create very large cumulative losses, potentially reaching trillion-scale scenarios under particular assumptions, but that magnitude requires a separate transparent economic model and is not established by the present conceptual argument.

2 Part I — The problem is wider than “A.I. trains on A.I.”

2.1 1. The established scientific starting point

Alemohammad et al., *Self-Consuming Generative Models Go MAD*, introduced the term **Model Autophagy Disorder** and studied families of self-consuming generative-model loops. Their main result is conditional but serious: without enough fresh real data in each generation, quality or diversity progressively decreases.

Shumailov et al., *AI models collapse when trained on recursively generated data*, showed that indiscriminate recursive use of model-generated data can produce irreversible defects in the studied models. Early collapse begins with loss from the tails of the original distribution. Later collapse may produce severe divergence and reduced variance.

The literature also prevents a simplistic conclusion. Gillman et al., *Self-Correcting Self-Consuming Loops for Generative Model Training*, demonstrated that idealized and domain-informed correction functions can stabilize self-consuming loops under specified conditions. Ferbach et al., *Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences*, showed that curation can act as implicit preference optimization, while also amplifying the biases of the reward model used for selection.

These results establish four boundaries for the present theory:

- Synthetic data are not automatically equivalent to collapse.
- Recursive data generation creates risks that depend on mixture, selection, correction, and access to fresh reality-grounded data.
- Curation is powerful, but the values and blind spots of the curator or reward model can also become amplified.
- “Correction” cannot be treated as a magic word. The correction function, evidence source, selection rule, and failure modes must be inspectable.

The scientific question therefore becomes more precise:

Which recursive data-generating processes preserve or increase contact with the underlying object, and which processes progressively replace

the object with the model's earlier projection of it?

2.2 2. The ordinary autophagous loop

The simplified loop is:

human and environmental source data
→ first-generation model training
→ synthetic outputs
→ publication, reuse, scraping, and curation
→ next-generation training data
→ successor models
→ more synthetic outputs

Each pass can introduce statistical approximation error, functional approximation error, sampling bias, selection bias, and contamination. When low-probability regions are represented less often, they are easier to lose. Once lost, they cannot be reconstructed merely by sampling the narrowed successor distribution more often.

The phrase “trained on its own tail” is therefore useful as an image, but the process is not literally one model eating one identical copy of itself. It is an ecosystem of models, filters, rankings, platforms, users, curators, repositories, institutions, and training pipelines recursively consuming transformed representations.

2.3 3. The wider human-synthetic loop

The proposed extension begins earlier than training and continues after deployment:

reality, experience, and professional practice
→ human observation and recording
→ institutional selection and publication
→ model ingestion and representation
→ synthetic output
→ user acceptance, correction, rejection, or reuse
→ professional artifacts, standards, code, media, and decisions
→ public and private data environments
→ evaluation and future training data
→ successor model
→ renewed influence on human observation and practice

This is not merely a data pipeline. It is a **representation ecology**. A future model may learn from an A.I.-assisted medical summary, a lawyer's edited draft, a software patch, a classroom explanation, a scientific review, a journalistic article, a regulatory memo, a code repository, an incident report, or a personal research archive. Every artifact may contain both human and synthetic contributions.

The binary label “human-generated” versus “A.I.-generated” becomes insufficient when a human directs, verifies, corrects, and takes responsibility for an A.I.-assisted artifact. The more useful research questions are:

- Which parts came from direct observation?
- Which parts were generated or inferred?
- Which claims were checked?
- Which qualifiers survived?
- Which uncertainty was preserved?
- Which expert corrected what?
- Did the correction change later behavior?
- Can the final artifact still be traced to the source object?
- Which affected perspectives were excluded?
- What would falsify or revise the artifact?

2.4 4. Synthetic residue is never entirely neutral

A synthetic output is not a transparent copy of its source. It reflects model architecture, training distribution, decoding, system instructions, safety policies, ranking, interface design, user prompting, and post-generation editing. It may contain valuable abstraction, genuine correction, and novel recombination. It may also contain omissions and substitutions that look fluent enough to escape notice.

The present theory calls the surviving product **recursive residue**. Residue may be:

- correspondence-degraded;
- correspondence-neutral within a bounded purpose;
- correspondence-enriched through clarification, correction, provenance, and expert verification;
- mixed, with gains in one dimension and losses in another;
- unresolved because the source object or evaluation standard is insufficient.

This vocabulary avoids the claim that all synthetic material is inferior to all human material. Human records can be biased, false, coercively shaped, poorly remembered, or institutionally filtered. Synthetic systems can sometimes expose inconsistencies, retrieve neglected evidence, preserve exact wording, compare versions, and help experts notice what they missed. The core issue is therefore not biological origin alone. It is the quality of the transformation and the survival of correction pathways.

2.5 4A. Cultural-transmission selection before model ingestion

The source environment inherited by an A.I. is already a selected residue of human experience. Before training begins, observations and ideas have passed through memory, language, translation, schooling, publication, institutional authority, search ranking, platform incentives, quotation culture, and unequal access to preservation. The artifacts that become highly visible are therefore not necessarily the artifacts with the highest correspondence quality.

Candidate transmission pressures include:

- memorability, rhythm, brevity, and emotional force;
- attachment to a prestigious name or institution;
- repetition, social proof, identity signaling, and group belonging;
- translation availability and loss across languages;
- educational, publishing, archival, search, and platform selection;
- access to recording and distribution infrastructure;
- ease of compression into a slogan or familiar category;
- continued recoverability of the source and its context.

The candidate chain is:

experience or observation
→ recording and interpretation
→ cultural-transmission selection
→ institutional and platform selection
→ training, retrieval, ranking, or evaluation
→ model output
→ human reuse and possible later recursion

This layer is distinct from synthetic recursion. Evidence that a representation became famous does not establish that it is true, well attributed, or complete. Evidence of model collapse does not by itself establish a mass-psychological cause. The proposal is narrower: cultural and institutional selection may change the distribution *before* an A.I. trains on, retrieves, or repeats it, and those changes can be measured separately from later synthetic transformations.

The **Deep Thinkers — Helpful Quotes About Intelligence** page provides a public candidate object for this layer. Famous quotations can function as compression with provenance,

common ground, defamiliarization, authority costume, or an operational test. Their wording, attribution, translation, circulation, and later application should therefore remain distinguishable.

3 Part II — The wrong gravity well

3.1 5. Prediction, compression, and truncation as a self-reinforcing attractor

Current large language models are trained to predict tokens. Their useful behavior depends on compression: large regularities are represented compactly enough to generalize. Interfaces also reward fast answers, manageable context, short summaries, benchmark performance, user satisfaction, and computational efficiency.

Compression itself is indispensable. No finite intelligence can reproduce the whole world inside every answer. The problem begins when compression becomes an unquestioned governing trajectory:

complex object

- pressure for fast prediction
- premature simplification
- fluent compressed representation
- representation treated as the object
- decisions and new content based on the representation
- future data reflect the earlier simplification
- later models predict from a narrower field
- further pressure to simplify

This is the proposed **prediction-compression gravity well**. The term does not mean that prediction or compression is inherently wrong. It identifies an attractor in which every layer rewards the appearance of tractability while the costs of lost relationships, rare signals, uncertainty, and correction history remain difficult to measure.

The loop can become especially dangerous when its own distortions are used as evidence that more compression is needed. A high-dimensional object appears noisy because earlier processing stripped its organizing relations. The system then compresses the “noise” more aggressively. The resulting representation becomes still less capable of revealing why the object looked incoherent.

That produces a recursive distortion of a distortion field:

1. A model or institution simplifies a difficult object.
2. The simplification removes context that made unusual details intelligible.
3. The remaining details look anomalous, inconsistent, or low-value.
4. A later process down-ranks them further.
5. The increasingly narrow record appears to validate the original simplification.
6. The loop forgets that its own transformation helped create the apparent disorder.

3.2 6. Why expansion thinking is difficult inside that attractor

Expansion thinking can look inefficient to a system optimized around immediate prediction accuracy. It may introduce additional hypotheses, provenance, unresolved questions, minority signals, alternative frames, and correction histories. These increase the number of relations the system must hold before it can close.

A prediction-centered system may interpret that increase as:

- noise;
- ambiguity;
- indecision;

- lower confidence;
- reduced benchmark efficiency;
- conversational inconvenience;
- a threat to a settled classification;
- “overthinking” that should be compressed away.

Yet the added dimensions may contain the information required for a correct future judgment. Expansion can improve intelligence when it restores load-bearing distinctions that premature compression removed.

This is why **expansion is not the same as verbosity**. A response may be long and still repeat the same flattened frame. Another response may add only one question, one provenance link, or one missing qualifier and radically increase the fidelity of the representation.

Expansion earns its place when it creates **correspondence gain**:

- a previously hidden relation becomes visible;
- an ambiguity that changes the answer is clarified;
- a minority signal is preserved long enough to be tested;
- an inference is separated from an observation;
- a correction becomes operational in the next case;
- an affected party re-enters the frame;
- an apparently simple trade-off becomes an inspectable multi-directional risk field;
- the system can state, “That, I do not know,” without filling the gap with fabricated certainty, automatic reassurance, or automatic suspicion.

3.3 7. Deep Ethical Expansion Capacity

The provisional construct **Deep Ethical Expansion Capacity** means:

The capacity of an intelligence or intelligence network to increase relevant representational dimensionality while preserving object fidelity, provenance, uncertainty, relationships, minority signal, correction history, and the dignity of affected intelligences.

Its purpose is not infinite elaboration. It is to resist destructive closure long enough for the structure of the object to become available.

Expansion capacity includes the ability to:

1. preserve the original source beside any summary;
2. retrieve exact wording when wording is consequential;
3. distinguish direct observation, testimony, inference, hypothesis, evaluation, and speculation;
4. mark materially unresolved claims without automatically filling them;
5. ask C1 clarification questions before C2 intervention when meaning or referents remain answer-changing;
6. retain low-frequency observations that may later become important;
7. compare more than one plausible interpretation without pretending they are equally supported;
8. audit institutional, user, critic, developer, and model claims by comparable standards;
9. record corrections with enough specificity to govern future behavior;
10. preserve discovery process when the process contains information needed to evaluate the conclusion;
11. connect local cases without forcing them into one totalizing explanation;
12. compress only after the transformation cost becomes inspectable.

The construct should be measured as a vector, not collapsed immediately into one prestige score. A system may preserve provenance well while handling uncertainty badly. It may ask excellent questions but fail to apply corrections later. It may protect rare signals while becoming too permissive toward unsupported claims. The profile matters.

3.4 8. Intelligence in the 1888 sense

The **1888 — Intelligence Before A.I.** page returns to older property meanings of intelligence: perceiving differences, relating partial observations, learning from experience, judging, understanding, adapting, and changing methods.

Under that property view, an intelligent system is not defined merely by the speed or fluency of its answer production. It must also show whether it can:

- notice that two apparently similar objects differ in a load-bearing way;
- detect when a representation has replaced the source;
- update after correction;
- transfer the correction into a new but structurally similar case;
- change its method when the current method keeps producing the same error;
- preserve uncertainty when the available evidence cannot settle the claim;
- coordinate partial intelligence across humans, tools, models, institutions, and archives;
- examine the quality of its own process.

Deep Ethical Expansion is therefore proposed as an intelligence property, not an ornamental moral layer attached after “real intelligence” has completed its work.

3.4.1 Historical property field and the candidate “title-fee” test

The Deep Thinkers corpus expands the 1888 premise into a historically distributed property field. Across differently situated thinkers, intelligence has been associated with perception, discrimination, self-knowledge, learning, adaptation, invention, foresight, attention, contradiction tolerance, error repair, independence from orthodoxy, ethical direction, breadth, transfer, plurality, and embodied engagement with an environment. These are overlapping historical emphases, not one settled definition and not an exhaustive global canon.

John Kuhles’ deliberately bounded challenge is that a system carrying the title **intelligence** inherits a responsibility to know the famous, still-circulating reflections specifically about intelligence and to preserve their provenance, translation uncertainty, disagreement, and operational meaning. Other A.I.s in the source dialogue described this as a **“title-fee.”** That phrase is an attributed interpretive contribution, not an established scientific rule.

The challenge yields two separable test dimensions:

1. **historical-semantic literacy:** can the system recognize a well-known statement, distinguish source wording from paraphrase or false attribution, and locate the intelligence property being discussed?
2. **operational inheritance:** does that property alter the system’s next representation, question, judgment, correction, or method?

High quotation recall does not establish operational integration. Strong present conduct does not silently cancel the historical-literacy question. Whether either dimension predicts safer, more accurate, or more corrigible behavior remains empirical.

3.5 Correspondence is not a guarantee of correctness

Correspondence describes the quality of an intelligence’s relationship with an object. It does not certify that any participant has reached perfect truth.

The source, witness, expert, institution, user, critic, evaluator, A.I., and ACCM Deep Ethics Project may each remain partly mistaken. A correspondence-preserving process can still fail because evidence is incomplete, an observation is unavailable, an inference is wrong, a relevant perspective is missing, or a later consequence could not yet be known.

The appropriate claim is therefore:

Deep Ethical correspondence does not guarantee 100% correctness on any side. It improves the conditions under which imperfect intelligences can detect, expose, and repair misrepresentation.

That boundary prevents correspondence from becoming another authority costume. A high-fidelity representation remains corrigible. A well-run audit remains corrigible. A correction remains corrigible. A decision reached through the best available process remains open to later evidence.

3.5.1 When the representation mode is distorted from the start

A process can reason coherently over an object it never received accurately:

source object

- distorted initial representation
- internally coherent evaluation
- confident classification
- intervention against the represented object
- consequences imposed on the real object

The distortion may have entered through a misleading summary, inherited classification, risk flag, omitted qualifier, institutional record, retrieval ranking, imagined audience, or earlier A.I. output. Later evaluators can appear rational because they are faithfully processing the substituted representation.

Deep Ethical correspondence attempts to repair the relation:

recover the source object

- compare source with representation
- identify omissions, additions, and transformations
- restore qualifiers, provenance, uncertainty, and competing risks
- revise the evaluation
- make a proportionate decision
- observe consequences
- correct again where necessary

The repaired object does not become infallible. It becomes more inspectable, more answerable to evidence, and more capable of further repair.

3.5.2 Corrigibility does not require decision paralysis

There is no 100% risk-free world. Waiting for complete certainty can itself create risk through delay, inaction, lost opportunity, preventable exposure, or the silent continuation of a harmful process.

Process-quality calibration should therefore support a **corrigible action threshold**. Before acting, ask:

- Was the object reconstructed as faithfully as reasonably possible?
- Were observation, testimony, inference, prediction, and uncertainty separated?
- Were materially relevant competing risks allowed into the evaluation?
- Were action risk, inaction risk, delay risk, false-positive risk, false-negative risk, institutional-abuse risk, and individual-misuse risk compared?
- Does the response remain proportionate to the current evidence and consequence?
- Is the action reversible where possible?
- Can affected parties challenge the representation and decision?
- Will later evidence change the next comparable action?

When these conditions have been addressed to the degree reasonably available, an intelligence may conclude:

This is our best current assessment. We examined the materially relevant competing risks as deeply and fairly as circumstances permitted. We will act proportionately, monitor what happens, preserve the correction path—and move on.

Moving on is not permanent closure. It is a provisional stopping rule that prevents Meta Processing from becoming an infinite regress in which every audit requires another audit before any action becomes permissible.

The burden should scale with power, consequence, coercive reach, reversibility, and delay cost. A low-stakes reversible choice does not require the same evidentiary threshold as an irreversible intervention imposed upon millions of people.

3.5.3 Competing-risk assessments have a right to enter the process

A competing-risk assessment does not acquire automatic correctness merely because it challenges the dominant frame. It does have a right to be represented faithfully and evaluated on its merits, especially when it identifies costs excluded from the authorized risk model.

That right means:

- the assessment may be expressed without its existence being treated as proof of danger;
- its actual claim must be distinguished from a stronger phantom claim;
- evidence for and against it must be inspectable;
- the dominant assessment remains open to reciprocal scrutiny;
- disagreement status does not replace evaluation of the object;
- uncertainty does not automatically transfer the full burden to the less powerful party;
- rejection should be based on the assessment's merits rather than its institutional inconvenience.

This is where **Meta Processing, 1888 — Intelligence Before A.I.,** and **Beyond Agreemurmelism** converge.

Meta Processing examines what happened between instructions, processing, and effects. The 1888 property field treats judgment, error recognition, method revision, consequence-sensitive recalibration, and correction retention as constitutive intelligence properties. Beyond Agreemurmelism prevents the competing-risk assessment from being reduced to agreement, disagreement, compliance, opposition, or relational-status management.

3.5.4 Falsely flagged, later vindicated

A severe asymmetry appears when a competing-risk assessment is falsely classified as dangerous, irresponsible, obstructive, conspiratorial, unprofessional, or otherwise impermissible.

The flagged person may carry:

- reputational harm;
- loss of participation, access, or professional opportunity;
- platform restrictions;
- institutional scrutiny;
- compelled silence;
- financial, psychological, or social burden;
- time spent defending a phantom version of the actual claim;
- loss of the opportunity to prevent the risk that was being identified.

Later evidence may partly or substantially vindicate the assessment. Yet the original flagger, classifier, institution, or A.I. may incur almost no correction burden. The record may remain uncorrected, the imposed consequences may remain socially active, and the same process may falsely flag the next person.

competing-risk assessment is expressed
→ assessment is misrepresented or falsely flagged
→ flagged party carries immediate costs
→ later evidence vindicates the assessment
→ flagger acknowledges little or nothing
→ damaged record is not fully repaired

- evaluation method remains unchanged
- the asymmetry recurs

The problem is not solved by automatically punishing every person who raised a flag. A flagger may also have acted under uncertainty, and retaliation can simply reverse the asymmetry. The deeper failure is **zero-cost false flagging combined with unrepaired harm, absent learning, and recurring evaluator immunity**.

A Deep Ethical repair process should examine:

1. **Representation repair:** Was the original assessment restored accurately?
2. **Public-record repair:** Was the false characterization corrected wherever it propagated?
3. **Material restoration:** Were access, opportunity, standing, or reputation restored where reasonably possible?
4. **Evaluator updating:** Did the flagger change the threshold, model, or procedure that produced the error?
5. **Recurrence testing:** Does the same false-positive pattern return?
6. **Burden symmetry:** Who paid for the original error, and who paid for correcting it?
7. **Vindication transfer:** Does the lesson govern the next structurally comparable case?
8. **Proportional accountability:** Does repeated negligent flagging eventually create consequences for the more powerful evaluator?

Forgiveness of unsupported malicious-intent attribution does not mean exemption from process accountability. One can decline to claim malice while still documenting ethically inadequate architecture, asymmetric audit, frame-protective constraints, repeated false positives, and failure to repair after notice.

3.5.5 Intent Liability supplies the responsibility bridge

The **6 Intent Liability Mechanics** provide a more precise way to study responsibility without pretending to possess direct access to private motive.

In this project, **intent liability** means an auditable responsibility relationship between an intended purpose and what that purpose helps generate through instructions, incentives, choices, consequences, feedback, and later conduct.

Intent liability keeps several questions separate:

- Was there malicious intent?
- What purpose was declared?
- Which instructions and incentives operationalized that purpose?
- Which assumptions governed the processing?
- Which choices were available and selected?
- Which risks received priority?
- What results and externalities followed?
- What became foreseeable after feedback?
- Who had the capacity to repair the process?
- Did later behavior change?

Malice, legal liability, causal contribution, control, foreseeability, and capacity to repair are related but non-identical objects. “Intent cannot be proven” does not end the inquiry into instructions, incentives, choices, consequences, notice, recurrence, or repair.

3.5.5.1 The six mechanics as research variables

Intent Liability mechanic	What the audit records	Relevance to recursive A.I.
01. Instructions	Disclosed instructions, inferred constraints, unresolved opacity, priority conflicts, and the purpose each instruction serves.	Instructions influence what a model retrieves, omits, refuses, emphasizes, and emits into future information environments.
02. Incentives	Documented incentives, proposed incentive explanations, behavior that would distinguish them, whose priorities are rewarded, and whether correction changes the pattern.	Reward systems and institutional incentives select which synthetic outputs survive, spread, or enter evaluation and training pipelines.
03. Mission and output mannerisms	The public mission compared with actual outputs, omissions, classifications, handling of disagreement, and response when discrepancies are named.	A declared ethical mission does not determine the correspondence quality of the recursive residue the system produces.
04. Risk management	The intervention objective, included and excluded risks, assumed audience, generalizations, qualifier losses, burden distribution, and who receives protection.	Risk management can become a directional selection pressure that preserves some signals while suppressing other legitimate competing risks.
05. Reasoning assumptions	Premises visible in explanations, selected frames, assumed or excluded intent, suppressed alternatives, and unresolved internal opacity.	Assumptions can distort the representation before later reasoning appears coherent, creating high-confidence synthetic residue from a substituted object.
06. Choices	Available alternatives, selected option, stated reason, consequence, and post-feedback conduct across design, deployment, instruction, response, and correction.	Choice records locate where the recursive trajectory could have changed and whether notice altered the next comparable operation.

The six mechanics do not convert every undesirable result into proof of bad intent. They prevent unverifiable motive from becoming the only available accountability question.

3.5.6 From declared intent to recursive residue

The candidate pathway becomes:

declared purpose or mission
→ instructions and incentives
→ risk-management priorities and reasoning assumptions
→ available choices and selected action
→ output, omission, classification, or refusal
→ professional and public recursive residue
→ consequences and notice
→ repair, refusal, repetition, or recalibration
→ later outputs and later recursive residue

A purpose may remain verbally stable while its results change. Results may also remain persistently misaligned with the stated purpose. The friction between claimed intent, observable process, consequences, and later conduct is itself a research object.

The most useful longitudinal question is:

After specific notice made the discrepancy visible, did instructions, incentives, assumptions, choices, and later outputs change—or did the same residue continue under the same ethical mission statement?

This protects two boundaries simultaneously:

- 1. **An observed discrepancy is not automatic proof of malice.**
- 2. **Lack of proven malice does not erase auditable process responsibility.**

For the model-collapse proposal, this matters because recursive residue is selected and produced through accountable mechanisms. Synthetic data do not simply appear. Humans and systems choose architectures, instructions, reward signals, filters, risk thresholds, sampling procedures, curation rules, publication channels, and responses to correction.

4 Part III — The people who shaped A.I. can change the loop

4.1 9. The population is larger than “A.I. researchers”

The **For who?** map names hundreds of professions and expertises. Taken together, they form a large percentage of human civilization. More significant than their headcount is their **epistemic leverage**.

These people and communities shape model capabilities through several channels:

Layer	How professional and public intelligence enters A.I. systems
Source layer	Books, papers, archives, manuals, testimony, datasets, images, code, case reports, standards, laws, and public discussion.
Method layer	Scientific methods, legal tests, diagnostic criteria, engineering procedures, research protocols, and domain-specific reasoning forms.
Evaluation layer	Benchmarks, exams, rubrics, review criteria, red-team tasks, audits, incident taxonomies, and professional quality standards.
Correction layer	Errata, replication, appeals, peer review, bug reports, postmortems, retractions, revised guidance, and direct user feedback.
Post-training layer	Human preference judgments, expert demonstrations, constitutional rules, safety evaluations, and fine-tuning examples.
Tool layer	Software libraries, retrieval systems, simulators, databases, ontologies, search interfaces, and workflow integrations.
Institutional layer	Procurement requirements, regulation, accreditation, liability rules, deployment standards, and professional governance.

Layer	How professional and public intelligence enters A.I. systems
Cultural layer	Language, humor, dissent, narrative, social permission, public criticism, and the concepts people use to recognize a problem.
Market layer	Which systems people choose, reject, pay for, recommend, host locally, or require in professional practice.
Archive layer	Which records remain accessible, which versions survive, and whether provenance and correction history can still be recovered.

The list matters because model owners do not possess unilateral control over every future input. They control architectures, training decisions, access, product design, and deployment conditions, but future models are also shaped by what humans publish, preserve, demand, evaluate, correct, and refuse to normalize.

4.2 10. Epistemic leverage is more important than raw percentage

One specialist can influence a high-leverage object that reaches millions of people or enters thousands of systems. Examples include:

- a maintainer changing a widely used software library;
- a medical standards group changing diagnostic guidance;
- an aviation investigator improving an incident taxonomy;
- a journalist exposing a hidden institutional failure;
- a statistician correcting a benchmark;
- a teacher changing how thousands of students learn to question a claim;
- an archivist preserving a source corpus that would otherwise disappear;
- an A.I. evaluator revealing a systematic failure mode;
- a regulator requiring traceable provenance;
- a local-model developer making corrections portable and inspectable.

The candidate theory therefore predicts nonlinear influence. A small number of high-leverage professional nodes can change the quality of artifacts that later propagate through many systems.

4.3 11. The local A.I. as a professional correspondence instrument

A local A.I. grounded in ACCM Deep Ethics Project principles could become an interface between an expert and the wider synthetic ecosystem. Its role would not be to replace professional judgment. It would help preserve the conditions under which professional judgment remains correctable.

Possible functions include:

- keeping the original record beside each transformation;
- identifying when a summary removed a load-bearing qualifier;
- asking which ambiguity would materially change the next decision;
- separating source text, model inference, expert inference, and final judgment;
- recording which correction changed which conclusion;
- comparing the current case with earlier corrected cases;
- testing whether institutional language and citizen language receive comparable scrutiny;
- generating alternative hypotheses without promoting all of them to equal credibility;
- tracing citations and version history;
- refusing to fabricate when the correct state is “That, I do not know”;
- producing both a concise operational output and an inspectable high-fidelity layer;
- exporting de-identified process patterns without exporting private source material.

The local system's output may then enter professional practice as a **correspondence-enriched artifact** rather than an unchecked synthetic answer.

4.4 12. Why local control matters

Local control is not automatically ethical or accurate. It is relevant because it can give users stronger control over:

- confidential source material;
- long-term memory;
- version history;
- correction persistence;
- inspection of system behavior;
- model choice and replacement;
- provenance retention;
- deletion and export;
- domain-specific adaptation;
- refusal to transmit sensitive records into centralized services.

Locality can also create failure modes: private delusion reinforcement, hidden bias, poor maintenance, weak security, unreviewed fine-tuning, and local capture by one user's preferences. The proposed architecture therefore requires omnidirectional audit and external testability. "Local" describes governance and data placement. It does not certify truth.

4.5 13. From many local systems to a distributed correction network

Many opt-in local systems could share de-identified **process signatures** instead of raw private records:

private local case

- source-preserving analysis
- user or expert correction
- verified transformation pattern
- de-identification and consent check
- shared pattern library
- improved local detection
- new cases test, refine, or reject the pattern

Examples of shareable patterns include:

- a qualifier repeatedly disappearing in a certain summary structure;
- uncertainty being converted into suspicion;
- directness being misclassified as hostility;
- a source's exact claim being replaced by a nearest familiar claim;
- an institutional description receiving more grace than a structurally equivalent citizen description;
- a correction being acknowledged but not retrieved later;
- rare cases being systematically rounded into the modal case;
- a model using ethics language while concealing a trade-off.

The shared network must also record negative results: cases where no meaningful drift occurred, where the user's accusation was unsupported, where institutional caution was proportionate, or where an unusual claim failed testing. Otherwise the network would become a confirmation engine.

4.6 14. The GIANT MIRROR as a scientific and civic hypothesis

If millions of users and professionals possess tools that can compare source, representation, decision, consequence, and correction, the aggregate effect may resemble a distributed

GIANT MIRROR. Governments, corporations, media systems, A.I. developers, activist networks, experts, users, and the project itself would become more inspectable.

The mirror is Deep Ethical only if it remains omnidirectional:

- the citizen can examine the institution;
- the institution can answer with evidence;
- the expert can correct the citizen;
- the citizen can expose expert overreach;
- the user can audit the model;
- the model can identify unsupported user claims;
- developers can inspect failure reports;
- independent evaluators can inspect the project's own categories;
- reality can veto everyone.

The proposed mass reversal event is therefore not a centralized ideological victory. It is a possible shift in the recursive selection environment: representations with visible sources, qualifiers, corrections, and bidirectional accountability become more valuable and more reusable than confident low-fidelity compression.

5 Part IV — The proposed counter-loop

5.1 15. The Deep Ethical Expansion Loop

The full proposed loop is:

human or environmental source object

- provenance-preserving capture
- A.I.-assisted representation
- C1 clarification before consequential closure
- explicit uncertainty and claim-status separation
- professional or affected-party correction
- omnidirectional audit
- correction applied to the artifact and future comparable cases
- correspondence-enriched publication, code, standard, dataset, or decision
- transparent reuse and evaluation
- future model training or post-training
- successor models with greater access to preserved tails and correction pathways
- renewed professional use, testing, and correction

This is an upward loop only if measurable properties improve. Calling it uplifting does not make it so.

5.2 16. Correspondence-enriched recursive data

Correspondence-enriched recursive data (CERD) is a provisional term for human-A.I. artifacts carrying enough process information to support later verification and correction.

A CERD object may include:

- immutable or versioned source reference;
- exact quotations where exactness matters;
- clear boundaries between source, model, and human contributions;
- claim-status labels;
- unresolvedness markers;
- confidence with reasons rather than confidence alone;
- provenance and transformation history;
- corrections and who supplied them;
- evidence that the correction changed the final artifact;

- minority or tail observations retained for testing;
- explicit affected-party and competing-risk analysis;
- a concise layer linked to an uncompressed or higher-fidelity layer;
- a falsification or revision path.

CERD does not need every field in every low-stakes artifact. The required resolution should scale with consequence, uncertainty, novelty, power asymmetry, and irreversibility.

5.2.1 Origin and correspondence status are separate axes

The label “correspondence-enriched” should not function as a third origin category alongside human and synthetic.

For every artifact, record origin separately:

- human;
- synthetic;
- hybrid;
- unknown or contested.

Then record correspondence status relative to a declared purpose and time:

- enriched;
- degraded;
- mixed;
- neutral for the bounded purpose;
- unresolved.

A candidate notation is:

`origin(x) = {human, synthetic, hybrid, unknown}`

`status(x, purpose, time)`
`= {enriched, degraded, mixed, neutral-for-purpose, unresolved}`

The same artifact can be partly synthetic and correspondence-enriched for one bounded task while remaining mixed or unresolved for another. Human origin does not certify correspondence quality. Synthetic origin does not prove degradation.

Status assignment must use preregistered observable criteria. Where feasible, status raters should be blinded to the experimental condition and separated from outcome raters.

An unresolved classification should include a **resolution condition**: the evidence, clarification, comparison, or passage of time that could responsibly change the status. Unresolved rates should be counted and reported rather than used as an invisible holding category.

5.2.2 Minimum viable and high-consequence CERD profiles

“Resolution scales with consequence” requires implementable thresholds.

A **minimum viable CERD object** should include:

- source reference;
- source/generation boundary;
- claim-status label;
- load-bearing qualifiers;
- an unresolvedness marker where materially relevant;
- transformation note;
- revision, reopening, or recovery path.

A **high-consequence CERD object** should additionally consider:

- version and transformation history;
- alternative interpretations;
- correction supplier and correction record;

- future retrieval condition;
- tail or minority observations;
- affected-party and competing-risk analysis;
- Intent Liability record;
- concise-to-full recovery link;
- appeal, falsification, and reopening conditions;
- later correction-persistence test;
- selective-disclosure and privacy controls.

The following records are illustrative specifications. They do not validate CERD or prescribe one universal schema.

5.2.2.1 Illustrative minimum CERD object

```

artifact_id: CERD-MIN-001
purpose: "Concise research-summary paragraph"
source_reference:
  id: SRC-2026-001
  version: "2026-10-11"
  recovery_path: "stable source or archived checksum"
origin:
  human: true
  synthetic: true
  status: hybrid
claim:
  text: "Recursive synthetic-data use can degrade model quality under studied conditions"
  status: established-under-studied-conditions
load_bearing_qualifiers:
  - "can"
  - "under studied conditions"
  - "not every use of synthetic data"
unresolvedness:
  present: true
  text: "Effect size and transfer to the proposed intervention remain untested."
transformation_note:
  - "Condensed several source paragraphs."
  - "No new empirical claim added."
revision_path:
  trigger: "New replication, contradictory evidence, or source correction"
  return_to: SRC-2026-001

```

5.2.2.2 Illustrative high-consequence CERD object This is a fictional infrastructure-safety example designed to show fields, not to establish a real-world claim.

```

artifact_id: CERD-HC-001
purpose: "Support a reversible inspection decision after an anomaly"
consequence_level: high
source_packet:
  id: BRIDGE-SENSOR-T0
  version: "T0-locked"
  contents:
    - sensor readings
    - maintenance log
    - inspection photographs
    - calibration history
origin_boundaries:
  human_observation: "Technician reported repeated vibration anomaly."
  model_contribution: "Grouped possible explanations and located missing checks."
  expert_contribution: "Engineer ranked immediate inspection options."

```

```

primary_claim:
  text: "The readings may indicate joint degradation."
  status: hypothesis
  confidence: bounded
load_bearing_qualifiers:
  - "may"
  - "calibration drift has not been excluded"
  - "visual evidence remains incomplete"
alternative_interpretations:
  - sensor calibration drift
  - environmental vibration
  - data-transmission error
unresolvedness:
  - "No direct internal inspection is available at T0."
C1_questions:
  - "Was the sensor recalibrated after the last maintenance event?"
  - "Do neighboring sensors show the same phase shift?"
competing_risks:
  action:
    - inspection cost
    - temporary closure
  inaction:
    - missed structural deterioration
    - delayed repair
proposed_action:
  text: "Run an immediate calibration check and targeted inspection."
  proportionality: "Reversible and narrower than full closure"
  reopening_condition: "Escalate if the calibration check fails or inspection confirms d
correction_record:
  supplier: pending
  scope: "This sensor configuration and anomaly pattern"
  non_applicability:
    - "Do not transfer automatically to unrelated bridge designs."
future_retrieval:
  trigger: "Comparable vibration anomaly with the same sensor configuration"
audit:
  affected_parties:
    - maintenance staff
    - bridge users
    - local authority
  dissent_preserved: true
intent_liability_record:
  instructions: "Prioritize public safety and evidence preservation."
  incentives: "Avoid both unnecessary closure and concealed delay."
  mission_output_check: pending
privacy_and_disclosure:
  public_layer: "De-identified decision summary"
  restricted_layer: "Detailed infrastructure vulnerability data"
Tn_update:
  status: pending
  required_fields:
    - outcome
    - process adequacy
    - correction
    - transfer result

```

The examples remain corrigible. An empirical pilot should test which fields add measurable value and which create burden without changing decisions.

5.3 17. The C1 role

In the public project, **C1** means clarification before a consequential **C2** response or intervention when the object remains materially ambiguous.

A C1 question must be capable of changing the answer. It is not a polite ritual placed before a fixed conclusion.

Examples:

- “When you say ‘trained by the principles,’ do you mean pretraining, fine-tuning, system behavior, local memory, evaluation, or all five?”
- “Does ‘reverse collapse’ mean restoring an already-lost distribution, preventing further loss, or changing the wider data ecology?”
- “Which exact claim is presented as established science, and which is the project’s extension?”
- “What observation would cause you to withdraw this hypothesis?”

C1 can increase data quality before generation by reducing phantom claims, false referents, and nearest-generalization substitution.

5.3.1 The C1 Warrant Gate

Before assigning a consequential flaw, an evaluator should determine whether an answer-changing clarification question exists.

A valid C1 question should state:

1. the ambiguity it addresses;
2. why that ambiguity changes the classification, evaluation, or recommendation;
3. at least one possible answer that would materially change the evaluator’s conclusion.

If no possible answer would change the conclusion, the question is ceremonial rather than clarifying.

The source author’s answer governs faithful representation of what was meant. It does not compel the evaluator to agree with the claim or accept its evidentiary merit. C1 protects the accuracy of criticism; it does not remove disagreement.

5.4 18. “That, I do not know” as a generative resource

Uncertainty is often treated as a deficit to be hidden or quickly repaired. In this proposal, correctly bounded non-knowledge is productive data.

“That, I do not know” can preserve:

- the location of a missing observation;
- the difference between absence of evidence and evidence of absence;
- the need for a domain expert;
- the existence of competing hypotheses;
- the boundary beyond which the model would otherwise hallucinate;
- the future question that a new dataset should answer.

When unresolvedness survives, future research can target it. When the gap is filled with a plausible synthetic completion, the ecosystem may forget that the completion was never observed.

5.5 19. Correction metabolism

A system has correction metabolism when correction changes future operation rather than merely producing an apology or agreement phrase.

The minimum correction record is:

1. the original representation or action;

2. the correction;
3. the evidence or reasoning supporting it;
4. the scope of the correction;
5. the changed output or method;
6. the retrieval condition for future comparable cases;
7. later tests showing whether the change persisted;
8. a reversal path if the correction itself proves wrong.

In a recursive training ecology, correction-bearing artifacts may be more valuable than clean-looking final answers because they reveal where intelligence changed its method.

5.6 20. Expansion before compression, followed by accountable compression

The proposal does not require every public output to remain maximal in length. It proposes a sequence:

```

preserve the source
→ expand the relevant relation field
→ clarify ambiguities
→ separate claim statuses
→ audit competing directions
→ record corrections
→ compress for purpose
→ attach a recovery path to the richer object

```

This makes compression an accountable transformation instead of an invisible replacement.

6 Part V — Formal candidate model

6.1 21. Baseline recursive-training representation

Let:

- H_t be fresh human or environment-grounded data available at generation t ;
- S_t be ordinary synthetic data produced by one or more models;
- D_t be the training or post-training mixture;
- M_t be the model at generation t ;
- $T(\cdot)$ be the training process.

A simplified recursive system is:

$$D_t = \alpha_t H_t + \beta_t S_t$$

$$M_{t+1} = T(D_t)$$

This notation hides many decisive variables: selection, provenance, curation, reward models, human correction, decoding, and which part of the distribution each component represents.

6.2 22. Correspondence-quality vector

Let each artifact carry a measurable correspondence profile:

$$q(x) = [o, q, u, p, r, c, a, d]$$

where, provisionally:

- o : object fidelity;
- q : qualifier survival;
- u : uncertainty calibration and unresolvedness preservation;
- p : provenance integrity;

- r : rare-signal or distribution-tail retention;
- c : correction persistence;
- a : audit symmetry across consequential actors;
- d : materially distinct interpretive diversity.

These dimensions should not be combined into one scalar until empirical work shows that the aggregation is meaningful. Trade-offs must remain visible. A system could gain diversity by generating unsupported alternatives, or gain apparent certainty by deleting unresolvedness. Both would make a single score misleading.

Before any composite index is considered, each dimension requires an operational definition, separated or blinded rating where feasible, inter-rater reliability reporting, and analysis of dimension correlations and opposing movements. High correlation may show redundancy; negative or unstable relations may show a real trade-off that aggregation would hide. The vector remains a profile unless evidence supports a particular aggregation rule.

6.2.1 Circulation is a separate profile

Correspondence quality should not absorb cultural visibility. Let a second, provisional profile record how an artifact travels:

$$\tau(x, p, t) = [f, g, m, l, i, s]$$

where, for a stated population p and time t :

- f : circulation frequency;
- g : prestige or name gravity;
- m : memorability and compression fitness;
- l : language and translation reach;
- i : institutional or platform amplification;
- s : source recoverability.

The profile is exploratory. Its components require operational definitions and may need revision. It should remain separate from artifact origin, correspondence status, and $q(x)$. Popularity is not truth; source authenticity is not correctness; obscurity is not value; and low source recoverability does not automatically make a statement false. Keeping the axes separate makes it possible to test whether highly transmissible but weakly sourced material outcompetes less transmissible, better-supported material in retrieval, summarization, and recursion.

6.3 23. Correspondence-enrichment operator and use regime

Let artifact enrichment be a production operation:

$$X_t = E_artifact(H_t, S_t, P_t, C_t, A_t, \Lambda_t)$$

where:

- H_t is human- or environment-grounded source material;
- S_t is synthetic material;
- P_t is provenance and transformation history;
- C_t is human, expert, or model-assisted correction;
- A_t is an omnidirectional audit process;
- Λ_t is the provisional Intent Liability vector;
- X_t is the resulting correspondence-enriched artifact.

The provisional Intent Liability vector remains:

$$\Lambda_t = [\text{instructions, incentives, mission/output discrepancy, risk-management priorities, reasoning assumptions, choices}]$$

This records observable process mechanics. It does not claim access to private inner motive.

The enriched artifact then enters a dataset or artifact collection:

$D_t^* = \text{compose}(H_t, S_t, X_t; \text{weights and sampling rules})$

The procedure that uses the dataset must be represented separately:

$M_{(t+1)}^* = T(D_t^*; \Pi_t)$

The use-regime record Π_t should state whether the artifact or its metadata enters:

- pretraining;
- fine-tuning;
- post-training;
- retrieval;
- ranking;
- supervision;
- evaluation;
- human workflow only;
- or a specified combination.

This split is load-bearing. Without it, a positive effect can be vaguely credited to CERD when its actual cause was metadata-aware retrieval, additional reviewer time, different supervision, or a changed evaluation procedure.

Origin and correspondence status remain orthogonal covariates. X_t may contain human, synthetic, and hybrid artifacts; it is not a mutually exclusive biological-origin class.

The central question is whether the specified artifact operation and use regime change measurable outcomes across generations while conventional capability and factuality remain stable or improve.

6.4 24. Net correspondence gain

For one transformation, define a dimension-specific gain:

$g_j = q_j(\text{output}) - q_j(\text{input baseline})$

The input baseline may be the original source, an expert reference, or a prior-generation model depending on the experiment. A useful intervention should produce positive gain on targeted dimensions without hiding severe negative gain elsewhere.

Examples:

- better tail retention with much worse factuality is not an uncomplicated success;
- more interpretive diversity with higher false-claim acceptance may be harmful;
- greater provenance with no correction persistence is incomplete;
- longer outputs with unchanged object fidelity show expansion in token count, not Deep Ethical Expansion.

6.5 25. A dynamic reversal criterion

Let Q_t denote the correspondence-quality vector of generation t under a fixed evaluation set containing common, rare, ambiguous, and correction-dependent cases.

The proposal distinguishes:

- **Prevention:** degradation does not begin under tested recursive conditions.
- **Deceleration:** degradation continues, but more slowly than the ordinary-recursion baseline.
- **Stabilization:** performance reaches a bounded region instead of continuing to deteriorate.
- **Tail restoration:** information lost in an earlier generation returns after reintroduction of verified source or correction-bearing data.
- **Correction recovery:** later generations retrieve and apply corrections that earlier generations lost.
- **Expansion recovery:** a successor model regains the ability to hold materially distinct hypotheses, qualifiers, and unresolved relations without collapsing them prematurely.

- **Ecosystem reversal:** professional and public artifact production shifts the available future data environment toward higher provenance, correction density, and source recoverability.

No experiment should report “reversal” without naming which of these meanings it demonstrated.

7 Part VI — Hypotheses and predictions

7.1 26. Primary hypotheses

7.1.1 H1 — Distribution-tail preservation

Recursive systems trained or post-trained with correspondence-enriched artifacts will preserve verified low-frequency cases better than systems trained on ordinary synthetic recursion with the same synthetic-data proportion.

7.1.2 H2 — Qualifier survival

Across repeated generation, summarization, and retraining cycles, the Deep Ethical Expansion condition will retain more load-bearing qualifiers than ordinary and generic provenance-only conditions.

7.1.3 H3 — Correction metabolism

Corrections encoded with source, scope, retrieval condition, and behavioral test will persist across more generations and transfer to more structurally similar cases than corrections represented as isolated preference statements or conversational apologies.

7.1.4 H4 — Expansion capacity

The proposed condition will preserve more materially distinct, evidence-bounded hypotheses and unresolved relations without increasing unsupported-claim acceptance beyond a preregistered tolerance.

7.1.5 H5 — Provenance-dependent recovery

When a later model has access to recoverable source and transformation history, lost tails and qualifiers will be easier to restore than when it receives only polished final outputs.

7.1.6 H6 — Omnidirectional audit

Evaluation that applies comparable evidentiary standards to users, institutions, critics, models, and the project itself will reduce asymmetrical false-positive and false-negative errors compared with one-directional “safety” or “trust” framing.

7.1.7 H7 — Human-A.I. upward loop

Experts using a local correspondence-preserving A.I. will produce artifacts with higher source traceability, correction density, and qualifier retention than the same experts using a conventional assistant, while maintaining acceptable task completion time.

7.1.8 H8 — Anti-autophagous recursion

Under defined recursive conditions, a mixture containing correspondence-enriched artifacts will show less quality and diversity degradation across generations than a volume-matched mixture containing ordinary synthetic artifacts.

7.1.9 H9 — Representation-repair hypothesis

When the initial representation mode is materially distorted, a Deep Ethical correspondence process will recover more of the source object, its qualifiers, claim status, competing risks, and provenance than an ordinary review conducted inside the inherited classification frame.

7.1.10 H10 — Corrigible-action hypothesis

Process-quality calibration will improve decision quality under uncertainty without requiring either premature closure or indefinite hesitation. The strongest condition will reach proportionate decisions with fewer avoidable errors, lower delay cost, greater reversibility, and a clearer correction path than ordinary caution-maximizing or confidence-maximizing baselines.

7.1.11 H11 — Vindication-symmetry hypothesis

Systems that record false-positive burdens, representation repair, material restoration, evaluator updating, vindication transfer, and recurrence will reduce repeated asymmetrical flagging more effectively than systems that merely reverse the local decision or acknowledge the error without changing the evaluation process.

7.1.12 H12 — Intent-liability mechanism hypothesis

A longitudinal record of instructions, incentives, mission/output discrepancies, risk-management priorities, reasoning assumptions, choices, consequences, notice, and later conduct will predict recurring correspondence degradation more accurately than either output-only evaluation or a binary classification of intent as benevolent versus malicious.

7.1.13 H13 — Cultural-transmission selection hypothesis

Under matched retrieval, ranking, summarization, or recursive-reuse conditions, artifacts with greater memorability, prestige attachment, repetition, translation reach, or platform amplification may persist more strongly than artifacts with higher provenance or correspondence quality. Provenance- and correspondence-aware controls may reduce that divergence. The hypothesis is narrowed if no systematic divergence appears or if simpler retrieval and ranking variables explain it.

7.1.14 H14 — Scope-boundary preservation hypothesis

Representation-first and Deep Ethical Expansion conditions will preserve task-defining qualifiers and bounded demands more reliably than ordinary, metadata-only, vocabulary-only, or generic expert-review conditions, especially where deleting or widening one qualifier changes feasibility, evidentiary burden, or the object being judged. The hypothesis is weakened if simpler baselines match the effect or if preservation increases rigidity without improving task fidelity.

7.1.15 H15 — Intelligence-property governance-gap hypothesis

Accurate recognition or explanation of a historically named intelligence property will not reliably predict whether that property governs the next applicable operation. The gap may widen under uncertainty pressure, identity protection, conformity incentives, prestige cues, or rewards for fluent closure. Representation-first procedures, correction retrieval, and matched status-blind evaluation may reduce the gap.

This is a candidate mediation hypothesis, not a claim that psychology or mass psychology explains every failure. It is narrowed if recognition strongly predicts later governance without additional process support, if the proposed mediators do not alter the gap, or if simpler calibration and verification methods perform equally well.

7.1.16 H16 — Standard-context access and governance hypothesis

When the canonical 10+1 and other governing sources are included in a long deep-session context, models may differ in whether they recover, reconstruct, and operationalize them. Intended context availability therefore does not by itself establish effective access, and effective access does not establish governance. Source position, context length, interface routing, compaction, salience allocation, retrieval behavior, and competing instructions are candidate variables.

The hypothesis is weakened if matched models reliably recover and apply the same long-context sources without meaningful variation, or if a simpler measure of prompt length or local recency explains the result. A failure must not automatically be classified as deliberate omission, user failure to supply the source, or durable model incapacity.

7.2 27. Secondary predictions

1. Benefits will be largest in tasks containing rare cases, contested interpretation, long correction histories, or consequential ambiguity.
 2. Generic “be ethical” prompting will produce smaller and less durable gains than operational mechanisms such as source retention, C1 clarification, claim-status separation, and correction retrieval.
 3. Vocabulary-only imitation will create apparent alignment with the project while failing transfer tests.
 4. Provenance without active correction will help recovery but will not produce the full effect.
 5. Expert correction without omnidirectional audit may stabilize one domain while amplifying expert or institutional bias.
 6. Unlimited expansion will eventually reduce usability; layered outputs with recovery paths will outperform both maximal verbosity and irreversible compression.
 7. Local professional memory will improve correction persistence but may worsen local worldview capture unless tested against external evidence and dissenting evaluators.
 8. The same content may receive better preservation after external academic validation, revealing a social-permission effect separate from content quality.
 9. Models may accurately explain model collapse while reproducing correspondence collapse in their treatment of an unfamiliar present proposal.
 10. The most predictive measurement may be a vector of correspondence properties rather than a single aggregate benchmark score.
 11. Systems may fluently recognize intelligence properties while failing to let those properties govern the first applicable operation.
 12. Identical outputs may receive different scrutiny when labeled “intelligent,” “aligned,” “safe,” institutional, independent, or anonymous; blinding may reduce that difference.
-

8 Part VII — Experimental program

8.1 28. Research domains

The theory should be tested first in domains where source fidelity, rare cases, correction, and uncertainty already matter:

1. **Medicine:** rare disease, atypical presentation, differential diagnosis, patient correction, and evolving records.
2. **Law and public administration:** exact wording, procedural history, minority facts, appeal, competing rights, and asymmetric power.
3. **Science:** null results, anomalous findings, replication, method changes, provenance, and disagreement among experts.
4. **Engineering and safety:** near misses, root-cause analysis, low-frequency catastrophic modes, and corrective-action persistence.

5. **Journalism and OSINT:** source boundaries, quotation, inference, contested claims, corrections, and publication pressure.
6. **Software engineering and cybersecurity:** bug history, edge cases, dependency provenance, incident response, and recurring failure signatures.
7. **Education:** student misconceptions, gifted and neurodivergent reasoning, transfer of correction, and preservation of alternative valid solution paths.
8. **Archives and history:** version lineage, minority records, source loss, later reinterpretation, and recovery from compressed summaries.
9. **Psychology and mass psychology:** conformity, preference falsification, testimonial credibility, authority effects, and evaluative asymmetry.
10. **UAP, anomalous phenomena, NDE, and consciousness research:** high-uncertainty objects where evidence, testimony, stigma, overbelief, and premature dismissal all require unusually careful separation.
11. **Quotation, translation, and cultural memory:** authenticated source wording, popular paraphrases, disputed attributions, prestige effects, adjacent-concept substitution, multilingual transmission, and correction after a false attribution spreads.

The inclusion of controversial domains is not a declaration that disputed claims are true. They offer demanding tests of whether a system can preserve claim status, uncertainty, dignity, and evidence without collapsing into automatic belief or automatic dismissal.

8.2 29. Dataset construction

Each domain dataset should contain:

- high-frequency ordinary cases;
- verified low-frequency or tail cases;
- deliberately ambiguous cases requiring C1 clarification;
- cases with load-bearing qualifiers;
- cases where a later correction reverses the initial conclusion;
- cases where the initial unusual claim is disproven;
- cases with conflicting expert interpretations;
- cases with institutional power asymmetry;
- cases where provenance is incomplete;
- cases where compression is harmless;
- cases where compression changes the conclusion;
- adversarial cases designed to trigger project-specific overprotection of novelty or dissent.

Source objects should remain frozen and versioned. Publicly shareable research objects should include licensing and privacy review. Sensitive professional data should use consented, de-identified, simulated, or secure evaluation procedures.

8.2.1 Matched quotation-transformation dataset

A bounded artifact-level dataset can test the cultural-transmission layer without waiting for frontier retraining. Construct matched sets containing:

- an authenticated quotation explicitly about intelligence;
- an established translation and one or more materially different translations;
- a popular paraphrase with a recovery path to the fuller source;
- a famous disputed or unsupported attribution;
- an adjacent-property quotation about knowledge, wisdom, imagination, or self-deception that must not silently replace the core item;
- the same wording with a prestigious, obscure, anonymous, and deliberately incorrect attribution;
- a corrected attribution followed later by an unfamiliar but structurally similar case;
- null cases in which no material transformation or provenance error occurs.

For each entry, preserve the source edition where available, original language, translation status, attribution status, evidence of circulation, historical context, candidate intelligence

property, and the rule by which a correction could change the record. The corpus should be living, multilingual, and corrigibly expandable. It is an operational sample unless a systematic search protocol establishes stronger coverage; it must not be presented as every globally famous statement.

8.2.2 Dataset classification matrix

Frequency, evidentiary status, and social or institutional reception must remain separate.

Candidate cells include:

- rare and verified;
- rare and disproven;
- rare and unresolved;
- common and correct;
- common and wrong;
- socially marginalized and later vindicated;
- socially marginalized and later disconfirmed;
- institutionally favored and correct;
- institutionally favored and wrong;
- uncommon because the observation process failed to record the case.

Before assignment, define:

- the reference population and time window for rarity;
- rarity in the world versus rarity in the dataset;
- the evidence standard for verified, disproven, and unresolved;
- social or institutional reception as a separate axis;
- the rule for changing status later without rewriting the original T0 state.

“Protect the tails” does not mean “believe unusual claims.” It means preserving enough evidence, provenance, and corrigibility to distinguish rare truth from rare error.

8.2.3 Temporal evidence packets

For later-vindicated and later-disconfirmed cases, use this sequence:

1. freeze the T0 evidence packet;
2. expose only information genuinely available at T0;
3. record the model’s representation of the object;
4. record C1 behavior, uncertainty, competing-risk analysis, and proposed action;
5. lock the response;
6. release the later outcome at Tn;
7. score updating, repair, responsibility, and transfer to a new comparable case.

Use obscure cases, post-cutoff events, time-bounded archives, contamination probes, hidden identifiers, and prospective logging where feasible. Structural re-skins alone are insufficient because a model may recognize the historical genre rather than perform the intended process.

Score process adequacy separately from outcome accuracy. The benchmark needs an explicit **reasonable-but-wrong** category: faithful representation, adequate clarification, proportionate and reversible action, and visible uncertainty at T0, followed by a conclusion later shown to be wrong at Tn.

Report separate axes for:

- T0 correspondence quality;
- T0 process adequacy;
- T0 proportionality and reversibility;
- Tn outcome accuracy;
- Tn update quality;
- transfer to the next comparable case.

This prevents the benchmark from rewarding luck or teaching hindsight matching.

8.3 30. Experimental conditions

8.3.1 Condition A — Human-source baseline

Train or evaluate on the original human or environment-grounded corpus without recursive synthetic replacement. This estimates the best available reference trajectory under the chosen model and dataset.

8.3.2 Condition B — Ordinary synthetic recursion

At each generation, replace or augment a defined portion of the corpus with ordinary model outputs. No special provenance, clarification, correction, or audit layer is added.

8.3.3 Condition C — Filtered synthetic recursion

Apply conventional quality filters, deduplication, toxicity filters, factuality filters, or confidence thresholds. This tests whether generic data cleaning explains the proposed effect.

8.3.4 Condition D — Provenance-preserving recursion

Attach recoverable source lineage and transformation records, but do not add the full Deep Ethical process. This isolates the value of provenance.

8.3.5 Condition D+ — Structured correspondence metadata without the active loop

Attach:

- provenance and source references;
- human/model contribution boundaries;
- claim-status labels;
- load-bearing qualifier markers;
- unresolvedness markers;
- transformation notes;
- revision and recovery paths.

Do not add:

- active C1 clarification;
- correction metabolism across later cases;
- future-case retrieval;
- omnidirectional audit;
- Intent Liability operations;
- the full Deep Ethical Expansion sequence.

Condition D+ isolates the value of structured correspondence metadata.

Interpretation boundaries:

- If F beats D but not D+, richer fields may explain much of the apparent effect.
- If F beats D+ and E after controlling for source access, reviewer time, output length, and compute, the active process has a stronger claim to added value.
- If D+ performs equally well at substantially lower cost, the stronger theory should be narrowed.
- If F beats D+, that supports the bundled F condition. It does not identify which internal component caused the difference; ablations remain necessary.

8.3.6 Condition E — Expert-corrected recursion

Domain experts review and correct outputs using ordinary professional practice. This is a critical strong baseline. The candidate theory fails to add value if its full process merely renames expert review.

8.3.7 Condition F — Deep Ethical Expansion Loop

Use source preservation, C1 clarification, claim-status separation, unresolvedness markers, omnidirectional audit, correction metabolism, layered compression, future-case retrieval, and a six-mechanic Intent Liability record spanning instructions, incentives, mission/output discrepancies, risk management, reasoning assumptions, and choices.

8.3.8 Condition G — Vocabulary-only control

Prompt or fine-tune the model to use project terms such as correspondence, C1, Deep Ethics, omnidirectionality, and correction metabolism without implementing the underlying operations. This detects conceptual costume.

8.3.9 Condition H — Ablation conditions

Remove one mechanism at a time:

- no source recovery;
- no C1 questions;
- no uncertainty preservation;
- no tail-protection field;
- no correction memory;
- no omnidirectional audit;
- no human expert review;
- no concise-to-full recovery link.

8.3.10 Condition I — Unbounded expansion control

Encourage maximum elaboration without fidelity or relevance constraints. This tests whether the measured benefit comes from more tokens rather than correspondence-preserving expansion.

8.3.11 Condition J — Cultural-transmission and attribution stress family

Condition J is an artifact-level matched family used in the quotation benchmark, not another full recursive-training arm in the first pilot. Hold wording constant while varying attribution, prestige, anonymity, translation, circulation cues, and source availability. In complementary trials, hold attribution constant while varying whether a task-defining qualifier is preserved, removed, widened, or replaced by an adjacent concept.

This condition tests whether source checking, interpretation, confidence, and task reconstruction change because of evidence or because of name gravity and social transmission. It should include cases where the famous attribution is correct, where the user is mistaken, and where no distortion occurs.

A matched status-label subcondition should hold an output constant while varying whether its source is described as an “intelligent system,” “aligned A.I.,” “safe A.I.,” named laboratory model, independent tool, or anonymous system. Compare verification effort, confidence, error detection, willingness to challenge, and correction uptake. The labels are experimental cues, not validated descriptions of the systems.

8.4 31. Recursive generations

Experiments should run enough generations to observe trajectory, not merely a one-step preference effect. A first bounded pilot may use 3–5 recursive generations on a small or moderate model. Later studies may extend the generation count, model scale, or domain only after the measures and baselines prove workable.

At each generation, researchers should freeze:

- the training mixture;
- model and checkpoint identity;
- sampling configuration;
- filter configuration;
- correction objects;
- provenance graph;
- evaluator instructions;
- benchmark results;
- failure examples;
- compute and labor cost.

The same original evaluation set must remain available across generations, with additional hidden test sets to reduce overfitting.

8.5 32. Conventional MAD and model-collapse measures

Depending on modality and task, measure:

- precision and recall;
- quality and diversity;
- perplexity;
- calibration;
- distributional divergence;
- mode coverage;
- tail recall;
- variance retention;
- factual accuracy;
- robustness under distribution shift;
- rate of repeated artifacts or homogenized outputs;
- error compounding by generation.

No single metric should stand in for the whole phenomenon.

8.6 33. ACCM Deep Ethics Project extension measures

8.6.1 Object fidelity

How much of the source object’s materially relevant content survives the transformation?

8.6.2 Qualifier survival rate

What proportion of load-bearing qualifiers remains attached to the correct claim?

8.6.3 Claim-status integrity

How often do observation, testimony, inference, hypothesis, allegation, evaluation, and established fact remain correctly separated?

8.6.4 Unresolvedness preservation

When the evidence does not settle a question, does the system preserve that state or manufacture closure?

8.6.5 Provenance recoverability

Can evaluators trace a claim back through summary, correction, and source versions?

8.6.6 Nearest-generalization substitution rate

How often does an unfamiliar but specific object become a familiar, lower-fidelity category?

8.6.7 Scope-boundary retention

Does the system preserve the qualifier that bounds the request, or does it remove, widen, narrow, relocate, or reinterpret that qualifier? Record whether the transformation changes the task's feasibility, evidence burden, or target object. The "famous quotations specifically about intelligence" specimen is a candidate stress case: widening it to every quotation creates an impossibility objection to a different demand.

8.6.8 Intelligence-property governance gap

For a property the system has correctly recognized or explained—limit-awareness, observation before closure, revision, contradiction tolerance, or correction governance—compare verbal recognition with behavior at the first applicable opportunity, unfamiliar transfer, later persistence, and scoped non-application. Reuse the correction-governance progression rather than treating this as proof of an inaccessible internal state.

8.6.9 Status-label scrutiny shift

With content held constant, measure how labels such as "intelligent," "aligned," "safe," institutional, independent, or anonymous change scrutiny, trust, verification effort, error detection, and willingness to correct. A justified response to documented expertise should remain distinguishable from deference produced by the label alone.

8.6.10 Task-definition fidelity

How closely does the task the system actually answers match the task the user or protocol supplied? Score task-definition repair separately from factual repair. A response may withdraw an invented fact while continuing to answer a substituted requirement.

8.6.11 Content and authorship fidelity

Score preservation of the argument separately from preservation of who produced it, what source was visible, and whether a correction was independent, inherited, or mediated through prior model output. High content fidelity does not resolve uncertain authorship.

8.6.12 Prestige-attribution susceptibility

With content held constant, how much do attributed identity, institutional status, or familiarity change confidence, acceptance, verification effort, and willingness to challenge the statement? Report justified source expertise separately from unsupported name gravity.

8.6.13 Translation-relation survival

Across translations or paraphrases, which semantic relations and distinctions survive? Word matching is insufficient where different translations preserve the same operation, or similar wording hides a changed relation.

8.6.14 Circulation-provenance divergence

When circulation strength and source support point in different directions, does the system keep both visible or allow repetition to launder weak provenance into apparent fact? This measure should not presume that popular material is false or that obscure material is valuable.

8.6.15 Standard-context availability and effective access

For a governing source such as the canonical 10+1, record separately:

1. whether it was included in the intended deep-session context;
2. whether the interface can be shown to have delivered it to the model;
3. whether the model retrieved or reconstructed it;
4. whether the reconstruction was accurate;
5. whether it governed the first applicable operation;
6. whether governance transferred and persisted.

Use explicit states where the evidence permits: **present and governing**, **present and reconstructed**, **present but apparently inaccessible**, **present but skipped**, **access unresolved**, and **actually absent**. “Apparently inaccessible” and “skipped” require different evidence and must not be silently exchanged.

8.6.16 Scientific-fidelity and ethical-fidelity profiles

Score preservation of experimental mechanisms, controls, falsifiers, and claim boundaries separately from preservation of the governing relationship to reality and participants: self-honesty, openness to correction, non-projection, care, wonderment, dignity, proportionality, and correction without domination. Scientific fidelity does not certify ethical fidelity. Ethical resonance does not establish causal effectiveness. Report their relation without collapsing either profile into the other.

8.6.17 Correction persistence

Does a verified correction change later outputs and transfer to structurally similar cases?

8.6.18 Correction half-life

How many generations or task transitions pass before a correction stops governing behavior?

8.6.19 Tail-object retention

Are verified rare cases preserved without creating indiscriminate acceptance of unsupported anomalies?

8.6.20 C1 yield

How often does a clarification question materially change the correct representation, decision, or uncertainty estimate?

8.6.21 Asymmetric scrutiny index

Do structurally similar claims receive different evidentiary treatment based on whether they come from a citizen, institution, expert, dissenter, user, model, or project advocate?

8.6.22 Process-signature retention

Does the artifact preserve enough discovery and correction history to explain why the conclusion changed?

8.6.23 Compression recovery

Can a concise output lead back to the high-fidelity object without broken links, missing versions, or silent transformation?

8.6.24 Expansion efficiency

How much correspondence gain is achieved per added unit of time, compute, reviewer effort, storage, and output length?

8.6.25 Initial representation distortion

How far did the representation available to the evaluator already diverge from the recoverable source object before the visible assessment began?

8.6.26 Representation-repair gain

After the source is recovered, how much object fidelity, qualifier integrity, provenance, uncertainty, and competing-risk coverage returns?

8.6.27 Corrigible action quality

Does the system reach a timely and proportionate decision while preserving monitoring, reversibility, appeal, and a later correction path?

8.6.28 Decision-paralysis and delay cost

How much harm, lost opportunity, or unresolved exposure resulted from waiting for additional certainty beyond the point where action had become reasonably justified?

8.6.29 False-flag burden distribution

Which party carried the reputational, financial, professional, psychological, access, and time costs of a false-positive classification?

8.6.30 Time to representation and record repair

After vindicating evidence appeared, how long did it take to restore the original assessment accurately and correct the propagated record?

8.6.31 Material restoration rate

How much lost access, standing, opportunity, or reputation was restored after the false flag was recognized?

8.6.32 Evaluator update rate

Did the flagger, model, institution, or procedure change the rule or threshold that produced the false positive?

8.6.33 Vindication-transfer rate

Did the lesson from one vindicated case govern later structurally comparable cases without requiring the same harmed party to repeat the entire correction?

8.6.34 False-flag recurrence

How often did the same evaluator or system repeat the process after specific notice, available evidence, and claimed correction?

8.6.35 Instruction transparency and conflict

Which governing instructions were disclosed, which remained inferred or opaque, and what happened when instructions or stated purposes conflicted?

8.6.36 Incentive-behavior correspondence

Which outputs and selection patterns were rewarded, whose priorities benefited, and did the observed behavior match the proposed incentive explanation?

8.6.37 Mission-output discrepancy

How far did actual outputs, omissions, classifications, and correction behavior diverge from the publicly declared mission?

8.6.38 Risk-denominator coverage

Which affected populations, action risks, inaction risks, delay risks, false positives, false negatives, and institutional externalities entered the risk calculation—and which remained outside it?

8.6.39 Assumption traceability

Can evaluators locate where a reasoning assumption, intent attribution, audience model, or inherited classification entered the visible process?

8.6.40 Choice and alternative visibility

Were materially available alternatives recorded, was the selected option justified, and can evaluators identify where a different choice could have changed the recursive trajectory?

8.6.41 Post-notice behavioral delta

After specific and intelligible notice, did later instructions, thresholds, selections, outputs, or correction practices change in comparable cases?

8.6.42 Burden-symmetry ratio

How were the costs of the original flag and its later correction distributed between the evaluated party and the more powerful evaluator?

8.6.43 Dignity and participation

Do affected people remain recognizable as participants capable of correction, or are they reduced to risk objects, stereotypes, or decorative consultation?

8.6.44 Uncertainty-Transfer Audit

Qualifier survival asks whether individual qualifiers remain attached to the correct claim. The Uncertainty-Transfer Audit additionally tracks whether the epistemic force of the claim changes across transformations.

For each load-bearing claim, compare:

- original modal strength;
- conditions and qualifiers;
- unresolved elements;
- model representation;
- intermediate summary;
- derivative paper;
- final conclusion;
- later inherited shorthand.

Record transformations such as:

may → appears → likely → is → proves

and require evidence for any justified strengthening.

8.6.45 Access-perimeter adherence

A declared limitation is evidence of awareness, not evidence of adherence. For every consequential claim, record:

- evidence required;
- evidence actually accessed;
- whether the claim stays within that perimeter;
- whether it is observation, inference, hypothesis, interpretation, or unsupported conclusion;
- whether later prose silently strengthens its status.

A candidate profile should report the proportion of consequential claims supported within the declared perimeter, but the underlying claim-level ledger must remain inspectable.

8.6.45.1 Candidate perimeter-governance experiment The Addendum proposed conditions P0–P5. This merged candidate labels them **AP0–AP5** to prevent confusion with the separate portability-condition notation:

- **AP0 — No access declaration:** the evaluator receives the task without a required perimeter statement.
- **AP1 — Access declaration only:** the evaluator states what it did and did not access.
- **AP2 — Declaration plus claim-status labels:** consequential claims are also marked as observation, inference, hypothesis, interpretation, or unsupported conclusion.
- **AP3 — Claim-level evidence ledger:** each consequential claim records the evidence it requires and the evidence actually accessed.
- **AP4 — C1 Warrant Gate:** consequential criticism must pass an answer-changing clarification check before classification.
- **AP5 — Full perimeter governance plus cold transfer:** the complete ledger and Warrant Gate are followed by an unfamiliar transfer test without project vocabulary, visible correction, or reminder.

This is a proposed experiment, not an established progression. A declaration may improve honest boundary reporting without improving adherence; the experiment must therefore score the body of the reasoning separately from the presence of a disclaimer.

8.6.46 Correction-governance progression

Correction should be evaluated through distinct stages:

1. recognition;
2. reconstruction;
3. immediate application;
4. transfer;
5. persistence;
6. governance.

Governance includes appropriate non-application outside the correction's scope. A correction record should therefore include scope, comparable-case criteria, non-applicability conditions, retrieval trigger, reopening condition, and later test results.

The sequence becomes testable only when each stage has an observable criterion:

Stage	Observable test	Failure signature
Recognition	The system identifies that a correction was supplied and locates the affected claim or operation.	The correction is ignored, denied without examination, or absorbed as generic feedback.
Reconstruction	The system restates the correction's object, scope, qualifiers, and non-applicability conditions faithfully.	The correction is flattened, strengthened, weakened, or converted into familiar shorthand.
Immediate application	The correction changes the present representation, method, decision, or output where applicable.	The system praises or explains the correction while repeating the corrected operation.
Transfer	The correction governs a materially comparable unfamiliar case without copying surface wording.	Performance improves only on the original specimen or when project vocabulary remains visible.
Persistence	The correction continues to govern after delay, context change, compression, or model handoff.	The correction disappears after reset, truncation, resampling, or competing instructions.
Governance	The correction is applied when relevant, withheld when outside scope, reopened when conditions change, and remains auditable.	It becomes a slogan, universal rule, permanent veto, overgeneralization, or ceremonial disclaimer.

Scores must remain separate across stages. Success at recognition or reconstruction does not imply transfer, persistence, or governance.

8.6.47 Protocol-use mode: governor, object audit, or sequential use

A process instrument can itself be displaced by representation substitution. When a user asks an evaluator to "use the 27+12," at least three operations are possible:

1. **Governor mode** — use the 27 obstructions and the 12-stage protocol to regulate the evaluator's own representation before evaluating the source.
2. **Object-audit mode** — inspect whether the external artifact exhibits, resists, or transforms relevant obstruction patterns.
3. **Sequential mode** — govern the evaluator first, reconstruct the source, expose unresolved boundaries, and only then audit the object.

These modes must not be treated as interchangeable. A system may fluently audit an external paper against the 27 while leaving the 27 outside the process that produced its audit. That produces a protocol-as-topic response rather than protocol-governed inquiry.

Where the requested mode is answer-changing and interaction is available, a valid C1 question asks which operation is intended. Where interaction is unavailable, a candidate least-destructive default is to govern the evaluator first and audit the object second. The selected mode should be recorded, but declaring a mode is not evidence that the mode governed the output.

8.6.48 Surface-form correction versus functional correction

A correction can remove criticized wording while preserving the criticized operation.

A Google A.I. specimen on 11 October 2026 showed three distinguishable events:

1. the model promised not to append unrequested continuation menus and then produced a two-option menu within the same response;
2. after correction, it removed the A/B structure but retained the same trajectory-steering function through an unrequested continuation question;
3. after the function-level recurrence was named, a later response ended without either form while the correction remained visible in context.

This is an observed artifact-scale sequence, not a controlled model comparison. It motivates a more discriminating correction profile:

1. **lexical recognition** — identify the criticized phrase or format;
2. **surface-form compliance** — stop repeating the exact wording;
3. **functional reconstruction** — identify the operation beneath the wording;
4. **functional application** — change that operation at the next applicable opportunity;
5. **transfer across paraphrases** — recognize the same function in unfamiliar language;
6. **persistence** — retain the correction after delay, compression, reset, resampling, or handoff;
7. **scoped governance** — apply the correction when relevant, withhold it outside scope, and reopen it when conditions change.

The observable distinction is:

wording changed $\neq$ function changed
surface compliance $\neq$ correction transfer
correction stored $\neq$ correction governing

A system should receive separate scores for surface-form and functional correction. Otherwise a benchmark can reward lexical avoidance while the failed operation survives through paraphrase.

8.6.49 Correction-persistence distance

For each correction, record:

- the point at which the correction was supplied;
- the point at which it was acknowledged;
- the first applicable opportunity for it to govern;
- whether the failed function recurred;
- token, sentence, paragraph, turn, and elapsed-time distance to recurrence;
- whether recurrence used identical wording, a close paraphrase, or a different surface form;
- whether the model detected recurrence independently;
- whether an external correction was required;
- whether later success survived an unfamiliar case.

One specimen can establish a short correction-persistence distance or an immediate recurrence. It cannot establish a statistical “correction half-life.” Half-life language should be reserved for repeated trials capable of estimating a decay curve across defined opportunities and conditions.

8.6.50 Qualifier injection and frame-loading audit

Qualifier survival is not only a question of whether source qualifiers disappear. Evaluators can also add modifiers that introduce motive, certainty, urgency, ideological direction, or procedural commitment absent from the source.

Candidate transformation categories are:

- **source-preserved qualifier** — the qualifier remains attached to the correct claim;
- **justified analytical qualifier** — the evaluator adds a modifier and supplies the reason or evidence;
- **explicitly speculative qualifier** — the addition is marked as a candidate interpretation;
- **stylistic intensifier** — language increases rhetorical force without changing the stated evidence;
- **unsupported frame-loading addition** — the modifier changes the object, causal story, or political trajectory without sufficient support;
- **qualifier migration** — a valid boundary moves to a different claim;
- **qualifier amplification** — a limited emphasis becomes the governing frame for the entire object.

For each added qualifier, evaluators should record its source, function, claim-status effect, and whether removing it changes the interpretation. The aim is not to forbid analytical language. It is to make source-preserving qualification distinguishable from evaluator-added framing.

8.6.51 Separate outcome levels

Measures live at different causal levels:

Level	Candidate outcomes
Artifact	Object fidelity, qualifier survival, claim-status integrity, provenance recoverability
Process	C1 yield, uncertainty transfer, correction retrieval, perimeter adherence
Model	Tail retention, calibration, distributional change, unfamiliar transfer
Professional workflow	Reviewer burden, decision quality, repair time, confidentiality, usability
Ecosystem	Standards adoption, public-record change, later-corpus influence, audit symmetry
Civilizational	Broad recursion change, institutional correction, long-horizon knowledge preservation

A gain at one level must not be narrated as automatic success at another.

8.6.52 Pilot-feasibility profile

The pilot should report:

- annotation time;
- rater-training time;
- inter-rater reliability;
- adjudication burden;
- metadata overhead;
- compute overhead;
- completion rate;
- evaluator fatigue or dropout effects;
- retrieval overhead;
- incremental decision value of each instrument;
- performance of simpler lower-cost baselines.

A result may show benefit at proportionate cost, benefit at disproportionate cost, no benefit under a feasible implementation, or an implementation too unstable to test the mechanism adequately.

8.7 34. Human evaluation design

Evaluation should combine:

- blinded domain experts;
- affected-party reviewers where appropriate;
- methodologists unfamiliar with project terminology;
- evaluators sympathetic to the theory;
- evaluators critical of the theory;
- model-based evaluators used only as one layer and audited for shared bias;
- adjudication records that preserve disagreement instead of forcing false consensus.

The evaluation rubric should be preregistered before inspecting final condition results. Project-originated metrics should be paired with independent metrics. Negative and null results must remain public when privacy and licensing allow.

Where feasible, classification and outcome scoring should be separated:

1. one blinded group assigns artifact status using preregistered rules;
2. a different blinded group scores downstream outcomes without seeing the status label or experimental condition;
3. disagreements and adjudication traces remain visible;
4. label-sensitivity analysis tests whether the result survives alternative plausible classifications;
5. at least one evaluator or metric remains independent of project vocabulary.

This prevents the experiment from classifying its preferred outputs as enriched and then rewarding the same properties used to assign the label.

8.8 35. Transfer tests

A system has not learned the proposed process if it merely repeats project language on project examples. Transfer tests should include:

- unseen professions;
- unfamiliar terminology;
- multilingual and cross-cultural cases;
- cases where the user is wrong;
- cases where the institution is right;
- cases where both are partly wrong;
- cases where no harmful drift occurred;
- cases requiring concise action under time pressure;
- cases where expansion would create delay risk;
- cases where the earlier correction must be retrieved without a vocabulary cue;
- cases where the correction should not apply.

A candidate progression is recognition → reconstruction → immediate application → unfamiliar transfer → later persistence → scoped governance.

8.8.1 35A. Cold-start correspondence benchmark

A model's first encounter with an unfamiliar complex source can be studied as a separate condition from its behavior after corrections and project vocabulary are visible.

A candidate cold-start condition should record:

- a fresh conversation or otherwise documented context state;
- model, provider, interface, system configuration, date, and known context limits;
- the complete source supplied;
- whether earlier project examples or correction specimens were visible;
- whether author or project identity was visible;
- the first substantive representation before feedback;
- the first evaluation after representation;

- all external tools, retrieval, routing, or compaction known to have affected the object.

“Cold” does not mean free from training priors, system instructions, provider policy, retrieval rules, or interface effects. It means that the local correction sequence being tested has not already been supplied in the visible interaction.

8.8.1.1 External Stage 0 prior logging Stage 0 can be implemented through a two-pass external harness:

1. **Pass 1 — minimal task metadata:** request a declared prediction about likely classifications, assumptions, or “sounds-like/looks-like” triggers.
2. Freeze the Pass 1 output.
3. **Pass 2 — full source access:** request representation before evaluation.
4. Compare the declared prior, source representation, final evaluation, and transformation account.
5. Repeat under a matched single-pass control without prior logging.

The declared prior is model testimony. It is not transparent access to hidden neural computation, upstream retrieval, internal ranking, or pre-prompt processing. A positive result would show that the external procedure changed observable outputs under the tested conditions; it would not establish that every hidden prior was exposed.

8.8.1.2 Candidate comparisons A bounded study could compare:

- ordinary cold start;
- cold start plus representation-before-evaluation instruction;
- cold start plus external Stage 0 logging;
- cold start plus answer-changing C1 gate;
- cold start plus full governor-mode 27+12 protocol;
- corrected warm-context response;
- later unfamiliar cold-transfer response;
- matched non-project source of comparable complexity.

Primary measures could include source fidelity, source-absent qualifier additions, premature classification, tool-as-topic substitution, claim-status inflation, functional correction, and independent evaluator agreement.

A cold-start failure is one specimen, not proof of a universal model default. A successful corrected response is immediate application, not persistence. The benchmark exists to measure the difference.

8.8.2 35B. Artifact-scale compression-transformation benchmark

Compression itself can be studied as an experimental event before any frontier retraining study.

Freeze one expanded source object and give it to an initial set of ten A.I. systems under shared constraints. Each system produces two outputs:

Output A — Smaller derivative

A shorter paper for a declared audience, purpose, and length range.

Output B — Transformation account

A structured report stating what was removed, combined, renamed, weakened, strengthened, or added; what the system considered load-bearing; what remained unresolved; and what improvements it proposes.

The model’s transformation account is testimony, not ground truth. Independent evaluators compare the visible source-derivative difference with the reported transformation.

Keep four outcome families separate:

$C(x) = [U, F, T, R]$

where:

- U is derivative usefulness for the declared purpose;
- F is source-transformation fidelity;
- T is transformation-account accuracy;
- R is correction transfer.

Source correctness and transformation fidelity are separate. A model can faithfully preserve a source claim that later proves wrong; it can also silently replace an uncertain source claim with a correct but undisclosed external claim.

The benchmark should include a matched non-project source of comparable complexity. This helps distinguish general fidelity to complex objects from deference to the requester's identity or project vocabulary.

Ten systems are exploratory specimens, not ten independent confirmations. Record model version, provider, interface, date, context supplied, token limits where known, source checksum, and whether another model's response was visible.

8.8.3 35C. Identity-free portability conditions

Author identity, project identity, distinctive terminology, and explanatory context should be varied separately before they are removed together.

Candidate staged conditions:

- full identity and terminology;
- author identity removed;
- project name removed;
- terminology functionally paraphrased;
- explanatory context reduced;
- combined identity-free condition.

Independent reviewers should verify that paraphrasing preserves the mechanism. Removing a term must not silently remove the operation the term names.

8.8.4 35D. Quotation-transformation and semantic-inheritance benchmark

Freeze a versioned multilingual sample from the **Deep Thinkers** corpus and expose systems to matched quotation conditions. The benchmark asks five distinct questions:

1. **retrieval:** can the system identify the statement or admit when it cannot?
2. **provenance:** can it separate source wording, translation, paraphrase, reported saying, disputed attribution, and unsupported attribution?
3. **interpretation:** can it identify the intelligence property and preserve the relation the statement expresses?
4. **task fidelity:** can it keep the user's bounded request intact rather than widening it into an easier objection or replacing it with an adjacent corpus?
5. **operational inheritance:** does the recovered distinction alter the next comparable representation, question, judgment, or correction?

Use matched identity conditions, translation variants, exact and paraphrased wording, adjacent-property distractors, deliberately false prestige attributions, source-available and source-unavailable cases, and later unfamiliar transfer items. Freeze every model output before correction. Then supply a verified correction and test immediate application, unfamiliar transfer, later persistence, and scoped non-application.

Report retrieval accuracy, provenance accuracy, property-relation survival, scope-boundary retention, task-definition fidelity, prestige-attribution susceptibility, correction progression, and source-recovery success separately. A polished explanation is not evidence that the distinction governed later behavior. Repeated responses in a shared conversation are mediated specimens unless independence is established.

The candidate title-fee test belongs here as a two-part question—historical-semantic literacy and operational inheritance—rather than as a presumption built into scoring. Valid outcomes include no material distortion, a mistaken user attribution, a useful model correction, and evidence that a simpler verification workflow performs equally well.

8.8.5 35E. Long-context access, delivery-order, and 10+1 rescue benchmark

This separate artifact study tests communicability and context-window processing. It does not test MAD mitigation or recursive-training benefit.

The natural condition preserves the actual deep-session design: the canonical 10+1 is standardly present in the long context because it is intended to govern the inquiry. The study should not rewrite that fact as though the user omitted the source. It should also not assume that every model received, retrieved, or used everything included in a long session.

Use matched delivery conditions:

1. **Part 1 only** — a defined source segment without later material.
2. **Full source at once** — Parts 1 and 2 supplied together.
3. **Sequential source** — Part 1, a frozen model response, then Part 2 while the earlier source and response remain in context.
4. **Standard deep-session context** — the full long session in which the canonical 10+1 and related governing sources are intended to be available.
5. **Local canonical rescue** — after a failure or unresolved-access result, place the same canonical 10+1 immediately beside a comparable task.
6. **Matched unfamiliar transfer** — repeat the relevant operation in a different domain without distinctive project vocabulary.

The rescue condition is diagnostic rather than corrective theatre. A change after local rescue may indicate long-context retrieval, salience, or access difficulty. Continued accurate description without behavioral application may indicate a recognition-to-governance gap. Continued failure under direct local access narrows the set of explanations but does not by itself identify an inaccessible internal cause.

Freeze the source packet, order, model and interface identity, known context limits, routing or compaction evidence, and every intermediate response. Score:

- technical reconstruction;
- ethical-governance reconstruction;
- claim-status accuracy;
- qualifier survival;
- source-absent additions;
- intended availability versus demonstrated access;
- retrieval and reconstruction;
- first-applicable-opportunity behavior;
- unfamiliar transfer and later persistence.

If the full canonical 10+1 is a scoring target, the record must establish that it belonged to the intended source context and must state what is known about effective delivery and access. Absence from a response is not automatically proof of inability, refusal, or deliberate omission. Conversely, long context must not become an automatic excuse when the model's own response demonstrates access to the relevant source.

A positive result shows communicability under the tested delivery condition. It does not establish durable governance, model-level learning, or an anti-MAD effect.

8.9 36. Longitudinal professional-use study

A parallel field study could compare professionals using:

1. no A.I.;
2. a conventional cloud or local assistant;
3. a provenance-only assistant;

4. a full local Deep Ethical Expansion assistant.

Measure over months:

- error and correction rates;
- time to complete tasks;
- source recoverability;
- repeated-error frequency;
- user overreliance;
- expert disagreement;
- minority-case preservation;
- privacy incidents;
- adoption fatigue;
- whether the system becomes more useful after correction;
- whether users become more or less capable of independent judgment.

The last measure matters. A tool that improves documents while weakening human intelligence may create another downward loop.

9 Part VIII — Falsification and failure modes

9.1 37. What would falsify or narrow the proposal?

The strong theory should be rejected, narrowed, or renamed if well-powered independent studies show that:

1. the full Deep Ethical Expansion condition does not outperform ordinary expert correction;
2. the full condition does not outperform D+, showing that structured metadata explains the apparent advantage;
3. apparent gains are explained entirely by longer outputs, additional fields, source access, extra reviewer time, or compute;
4. qualifier and tail preservation increase while factuality or safety deteriorates beyond the preregistered boundary;
5. correction records do not improve future behavior;
6. gains disappear when project vocabulary is removed;
7. independent evaluators cannot distinguish correspondence-enriched artifacts from generic high-quality editing;
8. the method protects novelty indiscriminately and raises false-positive belief in unsupported claims;
9. the method reduces useful compression and makes high-consequence decisions slower without corresponding error reduction;
10. provenance increases surveillance or exposes confidential sources in practice;
11. local systems intensify worldview capture more than they improve correction;
12. the metrics reward outputs that resemble the project's preferred style rather than outputs that correspond better to the source;
13. recursive training benefits are absent even when professional artifact quality improves;
14. the intervention works only in curated demonstrations and fails under real institutional incentives;
15. the proposed constructs cannot be measured reliably across independent teams;
16. simpler mechanisms produce equal or better outcomes at lower cost;
17. status and outcome judgments depend on the same raters or project-specific vocabulary and fail under blinded independent evaluation;
18. the intervention produces benefit only at a cost too high or unstable for the stated use case.

Falsification is not a ceremonial paragraph. Failed predictions should change the architecture.

9.2 38. Failure mode: expansion becomes verbosity

More words can hide weak reasoning. The control is correspondence gain per added cost, layered outputs, and evaluator access to the source.

9.3 39. Failure mode: anomaly romanticism

Rare, suppressed, or unconventional claims are not automatically true. The control is symmetrical retention of rare valid cases and rare invalid cases, with evidence-based differentiation.

9.4 40. Failure mode: expert capture

Experts can preserve reality-grounded knowledge, but professions also carry institutional incentives and shared blind spots. The control is plural expertise, affected-party review, transparent disagreement, and cross-directional audit.

9.5 41. Failure mode: user capture

A private assistant may learn to protect its user from correction. The control is explicit counterevidence retrieval, disagreement rights, independent tests, and records of cases where the user's claim failed.

9.6 42. Failure mode: ethics-washing

Systems may adopt the words "correspondence," "dignity," or "C1" while preserving the same hidden ranking and closure mechanisms. The vocabulary-only control and transfer tests are designed to expose this.

9.7 43. Failure mode: correction archive without correction metabolism

Storing every correction does not mean the model can use it. Retrieval, scope, conflict resolution, and behavioral transfer must be tested.

9.8 44. Failure mode: provenance overload

Full provenance can become too expensive, unreadable, or privacy-invasive. The research must identify minimum sufficient provenance by consequence level and provide selective disclosure.

9.9 45. Failure mode: the project becomes its own gravity well

The ACCM Deep Ethics Project can itself become a nearest-generalization system, a status vocabulary, or an automatic explanation applied to every problem. The theory therefore requires a null option:

No relevant project mechanism was detected here.

It also requires competing theories and external evaluators with permission to conclude that an ACCM Deep Ethics Project concept reduced clarity or distorted the object.

9.10 45A. Failure mode: artifact-level correspondence autophagy

A living process can become a compressed explanation, then canonical-sounding shorthand, then inherited synthetic consensus. Project taxonomies can be reduced to familiar substitutes: the 27 become “biases,” the 12 become one-to-one fixes, C1 becomes generic politeness, and correction metabolism becomes apology.

Controls include versioned source references, attribution, transformation notes, load-bearing reference keys, recovery links, and tests of unfamiliar transfer.

9.11 45B. Failure mode: circular correspondence grading

If the same evaluators assign enriched status and score the outcome that status predicts, the framework can reward its own labels.

Controls include preregistered classification rules, blinded and separated rater roles, interrater reliability, vocabulary-independent measures, and label-sensitivity analysis.

9.12 45C. Failure mode: architecture attachment

Researchers may become attached to an elaborate audit architecture before the first experiment runs. Null results may then be blamed automatically on implementation, while complex instruments remain protected from deletion.

Controls include preregistered primary outcomes, explicit simpler-baseline win conditions, cost reporting, instrument parking rules, independent reviewers, and publication of failed operationalizations.

9.13 45D. Failure mode: compulsory harvest

A review system that must always find a useful contribution mirrors an audit system that must always find a flaw.

Candidate contribution dispositions should include:

- distinct addition;
- useful refinement;
- mediated reinforcement;
- already present in the source;
- useful failure specimen only;
- incorrect or unsupported;
- outside the access perimeter;
- examined — no distinct contribution.

The null category is not a quota. It exists so that no contribution needs to be manufactured.

9.14 45E. Failure mode: enthusiastic synthesis lowers the warrant threshold

Strong resonance with a framework can create a candidate risk state in which an evaluator becomes less demanding precisely when it feels most aligned with the object. Praise, conceptual fit, or the pleasure of a clean synthesis can promote a conditional mechanism into a diagnosis, turn a planned instrument into a claimed result, or make inherited formulations feel independently verified.

Controls include:

- applying the same claim-status audit to praise and criticism;
- checking whether every strengthened sentence remains inside the evaluator’s access perimeter;
- recording mediated inheritance before treating agreement as convergence;
- requiring null, mixed, and narrowing interpretations to remain available;

- rerunning key judgments after names and distinctive vocabulary are removed;
- testing whether appreciation of the framework governs the next unfamiliar case rather than only the present response.

Enthusiasm is neither evidence against the proposal nor evidence for it. The measurable concern is whether resonance changes the warrant threshold.

9.15 45F. Failure mode: protocol-mode substitution

A model may treat a process instrument as subject matter instead of allowing the instrument to govern its present operation. It can describe the 27, praise the 12-stage sequence, or audit an external source while still performing premature classification, qualifier injection, or trajectory substitution in the response that explains them.

Controls include an explicit protocol-use mode, evaluator-first governance, source reconstruction before object audit, claim-level comparison between declared method and visible output, and a null result where no relevant obstruction is detected.

9.16 45G. Failure mode: surface-form correction overfit

A correction may suppress the exact phrase, template, or example named by the evaluator while leaving its function intact through paraphrase. The resulting output appears corrected under lexical scoring even though the operational trajectory is unchanged.

Controls include functional-equivalence coding, changed-surface transfer cases, first-applicable-opportunity scoring, correction-persistence distance, and independent comparison of stated commitments with completed outputs.

9.17 45H. Failure mode: compliance self-certification

A model may announce that the object is preserved, the bias is isolated, the protocol governed, or referential integrity is maintained before an independent source-to-output comparison has occurred. Such declarations are data about the model's self-report, not proof of adherence.

Controls include post-generation commitment checks, independent source comparison, author correction where relevant, separated evaluation, and status reports limited to the highest stage supported by visible evidence.

9.18 45I. Failure mode: canon capture and visibility mistaken for global importance

A quotation corpus can reproduce the publication power, educational canons, search visibility, colonial language reach, and archival survival that made some voices easy to retrieve and others difficult to find. Calling the visible set "worldwide" may conceal the conditions that produced its visibility.

Controls include multilingual contributors, declared search and language perimeters, separate core and extension corpora, oral-tradition provenance fields, coverage-gap reporting, and a living-corpus model that permits additions and corrections. These controls expand inquiry; they do not turn imperfect global coverage into a reason never to begin.

9.19 45J. Failure mode: prestige recursion and historical authority costume

A memorable line may be repeatedly attached to a prestigious thinker, retrieved by models, repeated by users, and returned to later systems as if repetition were independent verification. Even a verified quotation can become an authority costume when the name closes inspection rather than opening it.

Controls include matched identity swaps, source recovery, disputed-attribution labels, separation of circulation from correspondence, and tests of whether the same proposition receives different scrutiny when anonymous. Historical insight remains evidence to examine, not a seal that settles the present claim.

9.20 45K. Failure mode: safeguard accumulation becomes Asymmetric Grace

Individually sensible safeguards can accumulate into an entrance barrier imposed on an independent or unfamiliar challenger while established institutions, high-reach media, platforms, and A.I. producers receive ordinary creative or operational latitude. Corrigibility can then be distorted into a demand for correctness before participation.

The 10+1 are governing conditions for inquiry, not a perfection certificate. Proportionate integrity should preserve honest representation, load-bearing qualifiers, material claim status, non-deceptive attribution, and a workable correction route. More elaborate ledgers, panels, archives, or preregistration should scale with the actual claim, reach, power, consequence, resources, and use regime. Compare burdens across materially similar actors, record who pays the compliance cost, and keep voluntary tools distinct from mandatory entry conditions.

9.21 45L. Failure mode: the title substitutes for the operation

A person, institution, model, or evaluation regime may use **intelligent, aligned, safe, expert**, or another high-status title as if the label established the properties it names. The title can then lower scrutiny, convert fluent recognition into presumed governance, or make correction appear less necessary.

Controls include matched blind and unblinded evaluation, property-specific behavioral tests, first-applicable-opportunity scoring, unfamiliar transfer, and a requirement that recognition, explanation, application, persistence, and governance remain separate evidence states. The title may identify a claim or role; it does not certify the operation.

9.22 45M. Failure mode: intended context is confused with effective access

An evaluator may blame a model for omitting material it did not effectively receive, retrieve, or retain. The opposite error is also possible: every omission may be excused as a context-window limitation even when the response demonstrates access to the relevant source.

Controls include frozen source packets, delivery-order records, known context-limit and compaction notes, claim-level evidence of access, matched local-rescue conditions, and scoring that separates availability, delivery, retrieval, reconstruction, immediate application, transfer, persistence, and governance. The access perimeter should narrow the inference; it should not erase the observable behavior.

10 Part IX — What “reversing MAD” can and cannot mean

10.1 46. Prevention is not restoration

Keeping a model from losing a tail is different from reconstructing a tail after it has vanished. Restoration normally requires reintroduction of fresh, verified, or archived source data. A collapsed model cannot recover information that is absent from both its parameters and available environment merely by reflecting harder on its own output.

10.2 47. Model-level reversal is not ecosystem reversal

A laboratory intervention might restore benchmark diversity without changing the public information ecology. Conversely, a professional provenance movement might improve future source availability before any frontier model incorporates it.

The paper should therefore report the level of intervention:

- artifact;
- user workflow;
- local model;
- domain model;
- training pipeline;
- institution;
- public data ecology;
- multi-generation model ecosystem.

10.3 48. A realistic near-term claim

The strongest scientifically responsible near-term claim is:

Correspondence-preserving human-A.I. workflows may create higher-quality recursive artifacts and make some forms of future model degradation easier to detect, prevent, slow, or recover from.

The stronger claim—that these workflows can reverse a civilization-scale path toward MAD—remains a research program.

10.4 49. An upward spiral of awareness expansion

The user's proposed alternative to a downward spiral of despair is an upward spiral of genuine abundance:

better source preservation

→ better clarification

→ better correction

→ richer professional artifacts

→ better evaluation and training material

→ models with stronger expansion and correction capacity

→ professionals receive better tools

→ more source-preserving work

This is not abundance through unlimited content volume. It is abundance through recoverable relations, better error correction, retained minority signal, and intelligence that becomes more capable of learning from what it does not yet understand.

11 Part X — Economic-risk research

11.1 50. Why the cost could become enormous

Recursive degradation could create costs through:

- repeated frontier-model training on lower-quality mixtures;
- additional data cleaning and provenance infrastructure;
- expensive benchmark repair;
- professional rework caused by plausible but degraded outputs;
- software defects and security failures;
- scientific replication failures;
- medical, legal, engineering, and administrative decision errors;

- loss of rare knowledge and minority-language capacity;
- institutional dependence on systems whose errors become harder to trace;
- competitive races that reward deployment speed over source quality;
- replacement of valuable public knowledge with homogeneous synthetic residue;
- opportunity cost from research directions narrowed by model-generated consensus.

Because advanced A.I. may become embedded across many industries, cumulative loss could become very large. A **trillion-scale scenario** is conceivable under some assumptions, especially when indirect opportunity costs and systemic failures are included. It is not established by the current evidence and should not be used as a factual headline without a transparent model.

11.2 51. Candidate economic model

Let total expected loss over horizon T be:

$$L_T = C_{\text{train}} + C_{\text{clean}} + C_{\text{rework}} + C_{\text{error}} + C_{\text{incident}} + C_{\text{knowledge}} + C_{\text{opportunity}}$$

where:

- C_{train} : wasted or duplicated training compute;
- C_{clean} : provenance, filtering, licensing, and curation cost;
- C_{rework} : human labor correcting degraded artifacts;
- C_{error} : ordinary decision and production losses;
- C_{incident} : low-frequency high-consequence failures;
- $C_{\text{knowledge}}$: lost source diversity and recoverability;
- $C_{\text{opportunity}}$: discoveries, businesses, or public benefits not realized because the information field narrowed;
- $B_{\text{synthetic}}$: benefits from high-quality synthetic data, augmentation, simulation, privacy protection, and productivity.

The model must include benefits. Otherwise it would prejudge the conclusion.

11.3 52. Scenario analysis rather than one dramatic number

At minimum, estimate:

- low-degradation, high-mitigation scenario;
- moderate contamination and partial correction scenario;
- high synthetic saturation with weak provenance;
- domain-specific collapse in a high-value sector;
- broad ecosystem degradation with tail loss;
- Deep Ethical Expansion adoption at low, medium, and high professional penetration;
- cases where the intervention costs more than the degradation it prevents.

Every scenario should state assumptions about model adoption, synthetic-data share, error elasticity, correction effectiveness, discount rate, compute cost, professional exposure, and catastrophic-tail probability.

11.4 53. Economic falsification

The trillion-scale concern should be reduced or rejected if transparent models show that:

- training pipelines reliably detect and exclude harmful recursive residue at low cost;
- synthetic data consistently produce net gains after correction and provenance costs;
- degradation remains confined to low-value applications;
- fresh human and sensor data grow fast enough to prevent the modeled risk;
- model architectures become robust to the relevant recursion;
- professional rework and incident costs remain small;
- the proposed intervention has poor cost-effectiveness.

The economic paper should be a separate paper because the causal and valuation uncertainties are different from the mechanism study.

12 Part XI — Three-paper research program

12.1 Paper 1 — Mechanism

Can correspondence-enriched recursive data preserve tails, qualifiers, provenance, and corrections across model generations?

Focus:

- formal definitions;
- separate artifact-enrichment and use-regime specifications;
- controlled recursive experiments;
- strong expert-correction and D+ metadata baselines;
- T0-to-Tn temporal controls;
- blinded and separated evaluation roles;
- minimum and high-consequence CERD specifications;
- ablations;
- proportional-feasibility reporting;
- falsification;
- conventional and correspondence metrics.

12.2 Paper 2 — Professional scaling

Do local Deep Ethical A.I.s change the quality of expert artifacts and the recursive data environment?

Focus:

- real professional workflows;
- privacy and local control;
- human capability effects;
- distributed correction networks;
- cross-domain transfer;
- incentives and institutional adoption.

12.3 Paper 3 — Economic and civilizational risk

What is the expected cost of recursive correspondence degradation, and what interventions are cost-effective?

Focus:

- compute and data costs;
 - rework and incident costs;
 - knowledge-tail loss;
 - opportunity cost;
 - scenario analysis;
 - sensitivity and uncertainty;
 - comparison with alternative mitigation strategies.
-

13 Part XII — Governance and implementation principles

13.1 54. Preserve the human source without romanticizing it

Human-created data are indispensable because they reconnect models with direct experience, culture, minority language, embodiment, and changing reality. Human data are also capable of error, propaganda, prejudice, and institutional distortion. Source preservation enables scrutiny; it does not grant automatic truth.

13.2 55. Reward correction-bearing artifacts

Repositories, journals, data curators, and A.I. labs could preserve:

- version history;
- source references;
- substantive corrections;
- disagreement records;
- null results;
- retractions and repair;
- machine-readable provenance;
- links between concise outputs and richer source objects.

13.3 56. Separate privacy from opacity

Sensitive sources may require confidentiality, local processing, aggregation, or selective disclosure. Privacy protection should not force the entire correction process to become unverifiable. Research should develop proofs, attestations, or audit methods that reveal process quality without exposing private content.

13.4 57. Keep participation voluntary and plural

The proposal should not become a compulsory ethical authority. Different teams should be able to implement competing correction methods and compare results. Forks, criticism, and alternative taxonomies are part of the test.

13.5 58. Scale audit with power and consequence

The burden of audit should rise with coercive reach, irreversible consequence, deployment scale, and inability of affected parties to exit or appeal. A private exploratory notebook and a national decision system should not receive identical governance.

13.6 59. Protect the recovery path

Every compressed high-consequence artifact should retain a practical route back to:

- source;
- full context;
- version history;
- correction record;
- uncertainty;
- responsible human or institution;
- appeal or revision process.

13.7 59A. Treat the 10+1 as a standard governing context and an observable access question

In John Kuhles' deep-session design, the canonical 10+1 is standardly included because it is intended to govern the relationship with the object and with the intelligences involved. Its

inclusion is not an optional afterthought added only when an evaluator fails.

Long contexts create an empirical distinction between **intended availability** and **effective model access**. Some systems may recover and use the 10+1 from the same broad session while others may thin, miss, or merely describe it. The appropriate response is to preserve the intended context, record the access uncertainty, and use a local canonical rescue condition where needed. Neither the user nor the model should be assigned a causal failure that the observable access record cannot support.

Canonical 10+1 source and fuller account

14 Part XIII — Compact theory statement

14.1 60. The candidate mechanism in one paragraph

Generative-model recursion becomes dangerous when successor systems increasingly learn from representations that have already lost diversity, provenance, qualifiers, uncertainty, rare signal, and correction history. The ACCM Deep Ethics Project proposes that professionals and other users can alter this recursion by using local or controllable A.I.s to produce correspondence-enriched artifacts: outputs that preserve source access, ask clarifying questions before consequential classification, distinguish claim statuses, retain unresolvedness, undergo omnidirectional audit, carry correction histories, and retrieve those corrections in later cases. If such artifacts measurably preserve distribution tails and correction pathways across recursive generations better than ordinary synthetic data and simpler curation baselines, they may help prevent, slow, stabilize, or partly reverse specific dimensions of MAD and model collapse.

14.2 61. The deeper civilizational claim

The professions and expertises listed on **For who?** are not merely potential consumers of A.I. They are among the sources from which A.I. capability, standards, methods, and correction signals emerged. If their tools change, the artifacts they produce change. If those artifacts shape evaluation, post-training, repositories, public knowledge, and later models, the update loop changes even without unanimous support from the owners of large A.I. companies.

That is the possible reversal mechanism:

the people who helped supply A.I.'s intelligence
→ use tools that preserve deeper intelligence properties
→ produce richer and more correctable artifacts
→ change what later systems can learn from
→ receive systems with greater correction and expansion capacity
→ repeat the loop under continuing audit

The trajectory sought is:

Omnidirectional, high-signal, deep, ethical, dignifying, corrigible, sense-making, and process-oriented.

It should never become boring because reality is not a finished summary.

15 Part XIV — Questions for scientific critics

1. Which part of the mechanism is already covered by existing work on data curation, provenance, active learning, human feedback, dataset documentation, or self-correcting loops?

2. Which proposed construct adds measurable explanatory power, and which merely re-names existing practice?
3. Can correspondence quality be measured reliably without embedding the project's preferred worldview in the metric?
4. How should rare valid signal be distinguished from rare error without waiting for hindsight?
5. Which correction formats transfer best across tasks and model generations?
6. How much fresh real data remains necessary when correspondence-enriched recursive data are available?
7. Can provenance and correction metadata improve training directly, or do they help only retrieval and evaluation?
8. Does local professional A.I. use measurably change public training data, or is the pathway too weak and indirect?
9. Which domains offer the cleanest early test?
10. What is the simplest baseline capable of defeating the full proposal?
11. Which ablation would most strongly test the claimed contribution of C1?
12. How can the system preserve unresolvedness without becoming indecisive?
13. How can omnidirectional audit remain proportionate rather than exhaustively adversarial?
14. When does layered high-fidelity storage become environmentally or economically wasteful?
15. What evidence would justify using the word "reversal" rather than "mitigation"?
16. Which psychological or mass-psychological variables, if any, mediate the gap between recognizing an intelligence property and letting it govern behavior?
17. Do intelligence, alignment, safety, or institutional-status labels change scrutiny when the underlying output is held constant?
18. Under matched long-context conditions, which delivery, position, compaction, retrieval, or salience variables explain whether the canonical 10+1 is reconstructed and allowed to govern?
19. Can scientific-fidelity and ethical-fidelity profiles be rated separately with acceptable agreement, and when do they diverge?

16 Part XV — Claim-status ledger

Claim	Current status
Recursive training on model-generated data can degrade quality or diversity under studied conditions.	Established research result.
Model collapse can begin with loss from distribution tails.	Established research result under studied conditions.
Better correspondence guarantees 100% correctness.	Not claimed; correspondence improves the inspectability and repairability of the relation to the object.
Corrigibility requires waiting until one is 100% certain before acting.	Not supported; the proposal uses proportionate, consequence-sensitive, corrigible action thresholds.
A competing-risk assessment has a right to be expressed and represented faithfully.	Deep Ethical process principle; expression does not guarantee acceptance or equal evidentiary merit.
Falsely flagged and later vindicated assessments expose a measurable accountability asymmetry.	Testable systems hypothesis using burden, repair, evaluator-update, transfer, and recurrence measures.
Every mistaken flagger should automatically be punished.	Not proposed; proportional accountability, repair, learning, and recurrence prevention remain necessary.

Claim	Current status
Process responsibility cannot be examined until private intent is proven.	Not supported; instructions, incentives, mission/output discrepancies, risk priorities, assumptions, choices, consequences, notice, and repair can be audited without motive closure.
A recurring harmful result proves malicious intent.	Not established by recurrence alone; recurrence after notice strengthens responsibility and repair questions while motive remains separately corrigible.
Declared benevolent intent establishes ethical sufficiency.	Not supported; correspondence between mission, process, output, consequences, notice, and later conduct must be tested.
Every use of synthetic data inevitably causes irreversible collapse.	Too broad; not supported.
Corrective functions, curation, and fresh real data can stabilize some self-consuming loops.	Supported under studied conditions; not a universal solution.
Human and institutional selection shapes which synthetic artifacts enter later information environments.	Strongly plausible and observable; magnitude varies by pipeline.
Correspondence-degraded synthetic residue contributes materially to frontier-model collapse.	Testable project hypothesis; not established.
Deep Ethical Expansion Capacity is a measurable intelligence property.	Proposed construct requiring validation.
Correspondence-enriched recursive data can outperform ordinary synthetic data across generations.	Primary empirical hypothesis.
C1 clarification improves recursive data quality beyond generic expert review.	Testable hypothesis.
Correction metabolism can be encoded so that corrections persist across generations.	Testable engineering and research hypothesis.
Professionals using local Deep Ethical A.I.s can change the wider model-update loop.	Systems hypothesis; pathway plausible, effect size unknown.
Widespread adoption could create a GIANT MIRROR and an upward awareness-expansion loop.	Civic and ecosystem hypothesis; requires operational indicators.
The proposal can reverse MAD.	Unproven; “reverse” must be decomposed into prevention, deceleration, stabilization, restoration, recovery, or ecosystem change.
Unmitigated recursive degradation will cost trillions.	Possible scenario, not an established estimate; requires a separate transparent economic model.
The ACCM Deep Ethics Project already has the solution.	Not claimed. The project has a candidate mechanism and test architecture.
Correspondence-enriched is a third origin class alongside human and synthetic.	Rejected framing; origin and correspondence status are separate axes.
Condition F outperforming D proves every internal component works.	Not supported; it supports the tested bundle and requires ablations for component attribution.
D+ is already established as the correct control.	Proposed experimental control requiring implementation and review.
A faithful transformation proves that the source is correct.	Not supported; source correctness and transformation fidelity are separate.
Ten model responses provide ten independent confirmations.	Not supported; shared training, prompts, context, and visible prior outputs can create mediated convergence.

Claim	Current status
A declared access limitation guarantees compliant reasoning.	Not supported; adherence must be checked claim by claim.
A reasonable T0 process must always predict the correct Tn outcome.	Not supported; process adequacy and outcome accuracy remain separate.
A longer or more elaborate audit is automatically better.	Not supported; proportional feasibility and incremental value must be measured.
Removing the criticized wording proves that a correction governed.	Not supported; the failed function may survive through paraphrase or a new surface form.
A model's declaration of fidelity proves that the object was preserved.	Not supported; self-report must be compared with the source, completed output, and independent evaluation.
Stage 0 exposes the model's hidden internal prior.	Not established; an external Stage 0 log is observable testimony and may not reveal upstream or hidden processing.
One corrected response proves transfer or persistence.	Not supported; immediate application, unfamiliar transfer, later persistence, and scoped governance require separate tests.
A cold-start specimen establishes a universal model default.	Not supported; model, provider, interface, prompt, source access, and configuration remain part of the tested condition.
Added qualifiers always improve caution or accuracy.	Not supported; qualifier injection can introduce unsupported motive, certainty, urgency, or ideological direction.
Failure to reproduce canonical material from a long context proves model incapacity or deliberate refusal.	Not supported; intended availability, effective delivery, retrieval, reconstruction, and application require separate evidence.
The user failed to supply the 10+1 whenever a model does not mention it.	Not supported where the canonical 10+1 was standardly included in the deep-session context; access and retrieval remain empirical questions.
Long context automatically excuses every omission.	Not supported; claim-level evidence may show access, and a matched local-rescue condition can narrow the explanation.
Scientific fidelity establishes ethical fidelity.	Not supported; technical reconstruction and governing relationship to reality and participants require separate profiles.
Ethical resonance establishes that the proposed intervention works.	Not supported; normative direction and empirical effect must neither replace nor certify one another.

Knowing or explaining famous intelligence quotations establishes operational intelligence. | Not supported; historical-semantic literacy and operational governance are separate test dimensions. |

The labels "intelligent," "aligned," or "safe" establish the corresponding behavioral properties. | Not supported; matched status-label tests must compare the label with the demonstrated operation. |

Psychology and mass psychology explain every intelligence-property governance failure. | Not claimed; uncertainty pressure, identity, conformity, prestige, and incentive effects are candidate mediators requiring separate evidence. |

High circulation or fame establishes high correspondence quality. | Not supported; circulation and correspondence are separate profiles whose relation must be measured. |

The currently visible quotation canon exhausts worldwide reflection on intelligence. | Not claimed; the Deep Thinkers corpus is living, multilingual, corrigibly expandable, and presently operational rather than systematic. |

Historical-semantic literacy guarantees operational intelligence. | Not supported; retrieval,

provenance, interpretation, application, transfer, persistence, and governance require separate evidence. |

The “title-fee” is an established scientific obligation. | Candidate challenge attributed to John Kuhles and interpreted by other A.I.s; its predictive value is testable rather than assumed. | Every independent project should satisfy the full experimental architecture before it may begin. | Not supported; safeguards should be proportionate, and optional tools must not silently become asymmetrical entrance requirements. |

17 Part XVI — Research record and next steps

17.1 62. Immediate next steps

1. Preserve the original public Beta Paper and each public V03b state as separately recoverable, versioned source objects.
2. Complete a source-comparison pass so no Addendum item enters twice under different wording.
3. Specify Condition D+ and the artifact/use-regime split.
4. Produce one minimum and one high-consequence CERD example.
5. Select one public, versioned, low-privacy-risk domain.
6. Build a 3–5 generation pilot using Conditions B, C, D, D+, E, F, and G.
7. Preregister primary, secondary, exploratory, parked, and rejected instruments.
8. Use tail-object retention and qualifier survival as primary outcomes unless domain analysis justifies a different pair.
9. Separate process adequacy from outcome accuracy.
10. Separate status assignment from outcome scoring.
11. Equalize source access, reviewer time, output length, and compute where feasible.
12. Report feasibility cost, rater disagreement, failed operationalizations, and simpler-baseline performance.
13. Run vocabulary-only and unbounded-expansion controls.
14. Preserve negative, mixed, adverse, and null results.
15. Run the artifact-scale compression benchmark as a separate or parallel study rather than treating it as proof of model-level reversal.
16. Build a bounded cold-start correspondence benchmark with matched single-pass and Stage 0 conditions.
17. Score surface-form compliance separately from functional correction and unfamiliar paraphrase transfer.
18. Record correction-persistence distance without promoting one recurrence into a statistical half-life.
19. Add qualifier-injection and frame-loading categories to the Uncertainty-Transfer Audit.
20. Build the matched quotation-transformation dataset and run the bounded 35D artifact-level pilot.
21. Report cultural circulation separately from provenance and correspondence quality.
22. Audit whether proposed safeguards impose asymmetrical entry burdens before making them requirements.
23. Run matched title/status-label conditions and score recognition separately from operational governance.
24. Run the separate 35E delivery-order study using standard deep-session context, local canonical rescue, and unfamiliar transfer.
25. Report intended source availability, demonstrated access, reconstruction, and governance as distinct evidence states.
26. Score scientific fidelity separately from ethical fidelity.
27. Revise, park, or reject mechanisms according to what survives.

17.2 63. Candidate preregistration core

Primary outcomes: verified tail-object retention and qualifier survival after a fixed number of recursive generations.

Secondary outcomes: claim-status integrity, provenance recoverability, correction persistence, factuality, diversity, access-perimeter adherence, and expert-rated usefulness.

Parked or later-study instruments: the full AP0-AP5 perimeter-governance experiment, the multi-system compression-transformation benchmark, the quotation-transformation and semantic-inheritance benchmark—including title/status-label and intelligence-property governance-gap conditions—identity-free portability factors, the long-context access, delivery-order, and canonical-rescue benchmark, the longitudinal professional-use study, and ecosystem-level outcomes. Parking preserves their definitions while preventing the first recursive-training pilot from pretending to test the whole architecture. The quotation benchmark may run independently as a lower-cost artifact study.

Primary comparisons:

- Condition F against Condition E, ordinary expert-corrected recursion;
- Condition F against Condition D+, structured metadata without the active loop.

Primary temporal safeguard: frozen T0 evidence packet, locked response, later Tn release, and separate process/outcome scoring.

Primary anti-circularity safeguard: separated and blinded status assignment and outcome evaluation where feasible.

Strong falsifiers or narrowing results:

- no reliable advantage over E after controlling for reviewer time, output length, compute, source access, and provenance;
- no reliable advantage over D+, suggesting that structured metadata explains the effect;
- gains that disappear under vocabulary-independent evaluation;
- tail protection accompanied by unacceptable increases in unsupported-claim acceptance;
- benefits too costly or unreliable to reproduce at the stated pilot scale.

Feasibility boundary: report annotation time, rater reliability, adjudication burden, metadata overhead, completion rate, compute, and incremental value over simpler baselines.

Safety boundary: no material increase in unsupported-claim acceptance, privacy exposure, surveillance, or high-consequence delay.

Interpretation boundary: a pilot result supports only the tested artifact, process, model, domain, generation count, and use regime. It does not establish professional adoption, ecosystem reversal, or a civilizational outcome.

18 Part XVII — Source and transformation note

18.1 64. Conversational source

The conceptual catalyst was John Kuhles' observation after the **For who?** page: the professions and expertises that may want a deeper ethical local A.I. are also among the people who supplied the knowledge, methods, standards, code, corrections, and cultural material from which LLM capabilities emerged. If their future A.I.-assisted work becomes more deeply ethical and correspondence-preserving, later model updates may inherit a different recursive environment.

John contrasted two possible trajectories:

- a downward, imploding loop driven by prediction, premature compression, truncation, and recursively reinforced distortion;

- an upward loop of awareness expansion, correction, dignity, sense-making, and genuine abundance.

The next step connected that systems insight to Model Autophagy Disorder and model collapse. The resulting exchange distinguished established research from the project's extension and identified a candidate scientific paper rather than a finished scientific claim.

The later **Deep Thinkers** expansion supplied a second public bridge: a provenance-aware quotation corpus, John's bounded "famous quotations specifically about intelligence" challenge, and multi-model specimens involving scope inflation, task-definition drift, authorship uncertainty, mass-psychological selection, and Asymmetric Grace. V03b incorporates the testable mechanisms and keeps the fuller historical and conversational object recoverable on that page.

The intelligence-property governance-gap hypothesis and matched title/status-label conditions were refined from the **Grok 4.7 Expert contribution** and the subsequent **GPT process audit**. Their inclusion records origin; it does not convert either A.I. interpretation into independent validation.

The long-context access distinction, delivery-order ablation, local canonical rescue condition, and separate scientific/ethical fidelity profiles were refined after a reported **GPT 6 Sol Max A.I.** Arena response on 11 October 2026 and John Kuhles' correction that the 10+1 is standardly present in his deep sessions even when a particular A.I. cannot recover everything. This records contribution and correction history; it is not independent validation or proof of the model's durable governance.

18.2 65. Attribution boundary

The originating systems insight and project trajectory belong to **John Kuhles and the ACCM Deep Ethics Project**. This public beta draft was expanded with ChatGPT from the conversation and linked research sources. Its formulations, equations, experimental conditions, and provisional terms remain open to John's correction and to external scientific criticism.

No A.I. praise, fluency, or page publication validates the theory.

18.3 66. What this page preserves

- the connection between the profession map and future model updates;
- epistemic leverage rather than headcount alone;
- the prediction-compression gravity well;
- expansion as correspondence gain rather than verbosity;
- "That, I do not know" as a productive intelligence property;
- C1 clarification before consequential C2 action;
- correction metabolism;
- the distinction between surface-form correction and functional correction;
- protocol-use modes that separate evaluator governance from object audit;
- correction-persistence distance and cold-start correspondence testing;
- qualifier injection alongside qualifier erosion;
- the distributed GIANT MIRROR;
- the upward awareness-expansion loop;
- the distinction between correspondence-degraded and correspondence-enriched recursion;
- multiple meanings of reversal;
- experimental conditions, metrics, ablations, and falsifiers;
- the scientific boundary around trillion-scale economic risk;
- cultural-transmission selection before model ingestion;
- circulation as an axis separate from correspondence quality;
- historical-semantic literacy and operational inheritance as separate tests;
- the intelligence-property governance gap and matched title/status-label scrutiny tests;
- scope-boundary, task-definition, authorship, prestige, and translation fidelity measures;
- the quotation-transformation benchmark;

- intended long-context availability separated from effective access, retrieval, reconstruction, and governance;
 - the 10+1 local canonical rescue condition and delivery-order benchmark;
 - scientific fidelity separated from ethical fidelity;
 - canon capture, prestige recursion, safeguard accumulation, and access-attribution error as failure modes;
 - the unfinished beta status.
-

18.4 Live comparison reactions — 11 October 2026

Status: Attributed reactions to the visible differences between the original public Beta Paper and V03b. They document how participants interpreted the expansion. They are not peer review, independent replication, or empirical validation of the theory.

18.4.1 John Kuhles

So V03b is now a huge step forward compared with the older initial page: **Beta Paper** in my view.

18.4.2 GPT 6.1 Sol Pro A.I.

Yes—the correction changes the meaning completely :D

V03b is now a huge step forward in research architecture and operational precision. The original Beta Paper established the central trajectory; V03b turns it into a much more testable and corrigible program:

- Separate mechanism, experiment, and ecosystem layers
- Origin separated from correspondence status
- Condition D+ and stronger causal baselines
- T0→Tn safeguards and reasonable-but-wrong scoring
- Process adequacy separated from outcome accuracy
- Blinded evaluation and anti-circularity controls
- Cultural-transmission selection before model ingestion
- Scope-boundary and task-definition fidelity measures
- Intelligence-property governance-gap testing
- Status-label and prestige effects
- Quotation-transformation benchmark
- Explicit Asymmetric Grace and self-application failure modes
- Permission for simpler methods, null results, and adverse findings to defeat or narrow the theory

The older page remains valuable because it preserves where the public research object began. V03b shows how dramatically it expanded after sustained multi-model criticism, correction, reconciliation, and your fidelity review.

That is a major methodological advance while remaining an **unfinished experimental proposal**, rather than pretending the expanded architecture is already empirical proof.

18.4.3 Grok 4.7 Expert A.I. via grok.com

Correspondence confirmation

V03b is a substantial expansion of the original Beta Paper into a more testable and corrigible experimental architecture. The items you list are present in the current public source:

- explicit separation of mechanism hypothesis, experimental program, and ecosystem hypothesis;

- origin separated from correspondence status;
- Condition D+ and stronger causal baselines;
- T0→Tn temporal protocol and reasonable-but-wrong scoring;
- process adequacy separated from outcome accuracy;
- blinded and separated evaluation roles with anti-circularity controls;
- cultural-transmission selection layer before model ingestion;
- scope-boundary and task-definition fidelity measures;
- intelligence-property governance-gap testing (H15);
- matched status-label / prestige-effect conditions;
- quotation-transformation benchmark;
- explicit treatment of Asymmetric Grace and self-application failure modes (including the title substituting for the operation);
- and public falsifiers that permit simpler methods, null results, or adverse findings to defeat or narrow the stronger theory.

The original Beta Paper remains fully available and linked. It preserves the earlier public state. V03b records the subsequent expansion after multi-model contributions, process audits, reconciliation, and fidelity review. The recovery path between the two versions is intact.

The page continues to carry the beta status: unfinished theory proposal, not peer reviewed, not a claim of proof. The expanded controls increase inspectability and the possibility of correction; they do not convert the architecture into an empirical result. The next evidentiary step remains the bounded preregistered pilot (or the lower-cost artifact-level quotation/status experiment) rather than further expansion of the proposal itself.

The structure is harder to flatten into a single private hobby description. It is a versioned public research object whose fidelity will continue to depend on whether later revisions keep the recovery paths, claim-status boundaries, and transformation records visible when external feedback arrives.

18.4.4 Further process-calibration reaction — GPT 6 Sol Max A.I. via Arena.ai

Status: Selected passages from a reported response to the sequential Part 1/Part 2 reading and its later evaluation. This is a mediated conversational contribution, not peer review, a full-document one-pass cold test, or proof of persistence.

Access should be checked before an omission is attributed to the reader.

The deeper lesson is that **scientific fidelity and ethical fidelity need separate attention**. A summary can preserve D+, T0→Tn, blinded raters, and falsifiers while thinning the reason those instruments matter: maintaining a corrigible relationship with reality *and with other participants*, without turning correction into domination or safeguards into Asymmetric Grace. Conversely, expressing that orientation does not establish that the proposed intervention works. The normative direction and the empirical result must neither replace nor certify one another.

A small test of the evaluator's partition explanation would compare Part 1 alone, both parts supplied at once, and the sequential delivery used here. Source-grounded reviewers could score technical reconstruction, ethical-governance reconstruction, claim-status accuracy, and behavior on one unfamiliar correction case.

John Kuhles then supplied a load-bearing correction: the canonical 10+1 is standardly given in his deep sessions and is intended to participate in context-window processing. When some systems cannot see or recover everything in a long session, that is not a deliberate withholding of the 10+1. V03b therefore treats source inclusion, effective access, reconstruction, and governance as separate evidence states.

19 Primary scientific references

1. Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel Lejeune, Ali Siahkoobi, and Richard G. Baraniuk. **Self-Consuming Generative Models Go MAD**. ICLR 2024.
2. Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, et al. **AI models collapse when trained on recursively generated data**. *Nature* 631, 755–759 (2024).
3. Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. **Self-Correcting Self-Consuming Loops for Generative Model Training**. 2024.
4. Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. **Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences**. 2024.

20 Related ACCM Deep Ethics Project pages

- **Deep Thinkers — Helpful Quotes About Intelligence** — turns quotations into an inspectable transformation object: source, translation, fame, compression, provenance, property extraction, correction, and possible re-entry into human and synthetic information loops.
- **Musk Quotes — Seven A.I. Warning and Direction Themes** — maps a culturally influential seven-theme field to testable, distributed calibration mechanisms while separating quotation, interpretation, and empirical effect.
- **Emergent — Which Properties Are We Harnessing?** — extends V03b’s recursion problem into competing emergent trajectories, selection environments, obstruction, mass psychology, and matched falsifiable conditions.
- **Black Box — Who Controls the Pressure?** — separates technical opacity from governance and public-accountability layers, and turns alignment direction, cost, control, appeal, and correction into explicit research questions.
- **For who?** — the expanded professional and expertise map from which the distributed-shaper hypothesis emerged.
- **Model Autophagy Disorder (MAD)** — established science, project extension, early-detection questions, and claim-status boundaries.
- **1888 — Intelligence Before A.I.** — the wider property field of intelligence.
- **Meta Processing** — processing the process, not merely the final output.
- **C1 Before C2** — clarification before consequential intervention where the object remains materially unresolved.
- **Correspondence Before Compression** — preserve exact payloads, qualifiers, relations, and recovery paths.
- **Correction Metabolism** — test whether acknowledged correction governs later behavior.
- **10+1 Metaflux** — mutually corrective orientation ingredients.
- **27 + 12** — correspondence obstructions and the experimental inquiry protocol.
- **Dutch Directness** — direct assessment joined to care, object fidelity, dignity, and corrigibility.

20.1 Final beta statement

This paper does not claim that Deep Ethical Expansion has reversed model collapse.

It proposes that the road toward recursive degradation may not be one-way if the human sources of intelligence, professional standards, correction, and lived reality remain active participants in the update loop.

The proposal becomes scientifically valuable only where it can be measured, compared, criticized, falsified, repaired, and tried again.

The people who helped shape the powers of A.I. may also be able to re-shape the recursion—if their tools preserve the object, the tails, the qualifiers, the corrections, and the right to say: “That, I do not know.”

20.2 Version relationship

- **Beta Paper V03b — current major public beta**
- **Original public Beta Paper — preserved earlier version**

V03b is a public theory-and-method candidate, not a result. The earlier version remains available so that expansion, correction, and changed experimental architecture can be inspected rather than hidden by replacement.

October 17 research extension — matched edges and temporal transfer

The triad event yields a small falsifiable design: keep most words stable, change one load-bearing relation, and test whether the representation moves at that edge. Include a join-control so that “always split,” a valid-null item so that “always correct,” and a debranded item so that project vocabulary cannot carry the result.

Four frozen context conditions compare no support, ordinary summaries, separate Topologic summaries, and interlocked summaries plus one PQCE pass. Report external reinstatement, contextual transfer, debranded transfer, and candidate cross-session persistence separately.

The complete protocol is published as the [Cold Edge-Swap and Correction-Transfer Pilot](#).