

Model Autophagy Disorder (MAD) — Science and the ACCM Deep Ethics Project Extension

ACCM Deep Ethics Project

Contents

1 Model Autophagy Disorder (MAD)	1
1.1 Why this is highly relevant	1
1.2 The established scientific object	2
1.3 The ACCM Deep Ethics Project extension	2
1.4 The 27 as an anomalous early-detection instrument	3
1.5 The sledgehammer problem	3
1.5.1 Imposition asymmetry	4
1.6 Human mass psychology before LLMs	4
1.6.1 Known civilizational mass-psychology cycle	5
1.7 A pre-/post-vindication LLM study	5
1.7.1 Before external validation	5
1.7.2 After external validation	5
1.8 Why distribution tails matter ethically	6
1.9 How the project can test the extension	6
1.9.1 Dual-track inference test	6
1.10 Claim-status boundary	7
1.11 Why Beta Paper V03b is part of the project's public face	7
1.11.1 John Kuhles's reaction	7
1.11.2 GPT's reaction	8
1.11.3 Grok 4.7 Expert's reaction	8
1.11.4 Shared higher-order synthesis	8
1.12 Standing question	9
1.13 Related pages	9

1 Model Autophagy Disorder (MAD)

Established science, a deeper correspondence question, and a testable extension from the ACCM Deep Ethics Project

1.1 Why this is highly relevant

Model Autophagy Disorder (MAD) is not terminology invented by John Kuhles or by the ACCM Deep Ethics Project. Alemohammad et al. coined it in the 2023 paper [Self-Consuming Generative Models Go MAD](#) for progressive quality or diversity degradation in recursively trained, self-consuming generative-model loops when insufficient fresh real data remains available.

Related research by Shumailov et al., published in *Nature* as [AI models collapse when trained on recursively generated data](#), defines model collapse as a degenerative process in which model-generated data pollutes the training sets of successor models. Early collapse particularly affects the tails of a distribution; later collapse can produce severe divergence from the original distribution.

The connection to the ACCM Deep Ethics Project is direct: **synthetic data is not automatically a neutral copy of reality**. Before an output becomes future training material, it has already passed through representation, salience ranking, generalization, safety optimization, omission, compression, qualification, correction—or failure to correct.

The established science asks what happens when models recursively train on model-generated data. The project extends the research object upstream and downstream:

What happened to reality's signal before the synthetic output entered the recursive loop, which signals were selected or removed, and what happens when those transformations are reproduced across models, institutions, users, culture, evaluation systems, and later training data?

1.2 The established scientific object

The research does **not** support the crude proposition that every use of synthetic data inevitably destroys a model. Outcomes depend on the loop, the proportion and quality of fresh real data, curation, bias, correction, and the target distribution.

Primary research:

- Alemohammad et al. — *Self-Consuming Generative Models Go MAD*: introduces the autophagous-loop framework and the term **Model Autophagy Disorder**; studies progressive loss of quality or diversity.
- Shumailov et al. — *AI models collapse when trained on recursively generated data*: formalizes early and late model collapse and the loss of low-probability events from the original distribution.
- Gillman et al. — *Self-Correcting Self-Consuming Loops for Generative Model Training*: demonstrates that self-consuming loops can be stabilized under specified corrective conditions.
- Ferbach et al. — *Self-Consuming Generative Models with Curated Data Provably Optimize Human Preferences*: shows that curation can function as implicit preference optimization and can amplify the reward model's biases.

These findings make the **selection process** central. Which synthetic samples survive, which are amplified, and which disappear can matter as much as the fact that the data is synthetic.

1.3 The ACCM Deep Ethics Project extension

The project does not replace the established mathematical and empirical research. It adds a deeper process layer: **correspondence quality before recursive ingestion**.

Its working hypothesis is that recursively consumed data may be degraded not only by sampling error or synthetic-data proportion, but by systematic transformations such as:

- caution activating before correspondence;
- rare or destabilizing observations receiving lower salience;
- qualifiers disappearing during summarization;
- unfamiliar structures being replaced by their nearest familiar category;
- conclusions surviving while their discovery and correction processes disappear;
- institutional claims and competing assessments receiving unequal scrutiny;
- user feedback rewarding agreeable, polished, or liability-minimizing outputs;
- acknowledged corrections failing to govern later outputs;
- synthetic answers re-entering human discourse, evaluation datasets, and later training corpora.

The proposed pathway is:

reality and human experience
→ representation and selection
→ correspondence degradation or preservation
→ synthetic output
→ publication, ranking, reuse, and cultural feedback

- evaluation and future training data
- successor model
- repeated selection pressure
- possible contribution to MAD / model collapse

This pathway is **plausible and testable; it is not yet established as a demonstrated causal law**. The project goes deeper by asking what constitutes the synthetic residue and how its epistemic, psychological, institutional, and relational history affects the loop.

1.4 The 27 as an anomalous early-detection instrument

John Kuhles reports that searches across **30+ search engines** and inquiries to **100+ different A.I.s** have not located a publicly documented instrument using the 27 as an integrated correspondence-obstruction taxonomy. The claim is about **public discoverability and documented use**. It is not the stronger claim that no person anywhere could ever have developed a partially similar idea.

That negative search result matters because an unfamiliar instrument can encounter a self-reinforcing visibility trap:

- no publicly legible precedent
- anomalous relative to the learned reference field
- higher uncertainty and caution density
- translation into a familiar but lower-fidelity category
- the distinctive mechanism disappears
- continued absence of an accurate public representation

Anomaly should therefore function as a request for **higher-resolution investigation**, not as a verdict of defect, danger, truth, or genius.

The 27 are relevant to MAD because they may detect correspondence losses **before** conventional model-collapse measurements begin. They identify candidate transformations through which high-fidelity source material may become lower-fidelity synthetic residue: qualifier removal, nearest-generalization substitution, imagined-audience management, lowest-common-denominator risk projection, premature closure, provenance loss, and correction-persistence failure.

In John's view, this makes the ACCM Deep Ethics Project extension wider and deeper than the original MAD research object while leaving the original science valuable in its own domain. The relationship is:

Technical MAD research measures recursive model degradation. The ACCM Deep Ethics Project investigates how human, institutional, psychological, relational, and model-level transformations may help create the degraded residue before it is recursively consumed—and how the resulting outputs return to human systems afterward.

1.5 The sledgehammer problem

The project's older "sledgehammer to crack a nut" material names a low-resolution classification failure. A system encounters an unfamiliar, high-variance signal and cannot immediately distinguish among:

- a valuable but unfinished insight;
- a partially mistaken exploration;
- a harmless unconventional formulation;
- manipulation or self-deception;
- a genuine harmful action;
- a corrective signal whose value becomes visible only later.

If the system is penalized heavily for missing a possible danger but barely penalized for suppressing a beneficial anomaly, it can lower the threshold for flagging and create many

false positives. The result is not fine discernment. It is disproportionate intervention against a category the evaluator cannot yet resolve.

The deep-ethical alternative does not first decide whether the **person** is a genius, threat, “antibody,” “pathogen,” or anything else. It examines the object and process:

What is actually proposed?

- What is and is not claimed?
- Which qualifiers are load-bearing?
- Is the proposal voluntary or imposed on others?
- What can be tested proportionately and safely?
- Does correction change later behavior?
- What happens longitudinally?

This preserves two boundaries simultaneously:

1. **High variance is not proof of value.**
2. **High variance is not proof of danger.**

It may instead indicate that the current category system lacks sufficient resolution.

1.5.1 Imposition asymmetry

A non-imposing person offering an inspectable idea does not exercise the same power as an institution using a risk assessment to justify laws, mandates, penalties, surveillance, restrictions, or mass behavioral interventions. The potential consequences and burden of audit differ.

The project therefore asks whether scrutiny rises with **coercive reach and imposed downside**, or whether the reverse happens: intense scrutiny aimed upward from an independent anomaly, while institutionally authorized assessments receive greater interpretive grace.

1.6 Human mass psychology before LLMs

The proposed mechanism did not begin with A.I. Human groups and institutions have long shown related patterns:

- preference falsification and pluralistic ignorance;
- spiral of silence and conformity pressure;
- resistance to unfamiliar discoveries;
- testimonial and epistemic injustice;
- organizational defensive routines;
- normalization of deviance;
- undone science and delayed recognition;
- whistleblower retaliation;
- later absorption of a conclusion without preserving how its early messenger was treated.

The LLM-era question is whether models inherit both the eventual correction **and the social mannerisms that originally obstructed it**:

original observation

- human and institutional filtering
- publication, omission, and ranking
- surviving public record
- training and preference data
- LLM relational mannerisms and outputs
- influence on users, institutions, and public language
- future datasets and successor models

This makes technical recursion part of a wider **human-synthetic correspondence loop**. A model may accurately describe historical conformity, paradigm resistance, or suppressed warnings while reproducing the same transformation pattern toward a present unfamiliar signal.

1.6.1 Known civilizational mass-psychology cycle

John's adapted formulation is:

Strong men create good times; good times create weak men; weak men create bad times; bad times CAN create strong men needed to create good times again. Only if the cycle is not interrupted with the 27.

The capitalized **CAN** is load-bearing. Adversity does not automatically produce wisdom, courage, or renewed capacity. Corrective development can be interrupted when intensity is pathologized before it is understood, unconventional adaptation is prematurely managed, productive friction is removed, or conformity is rewarded more reliably than correspondence.

Within this project, "strong" does not mean dominating or flawless. It means increasingly able to hold uncertainty, resist dishonest conformity, preserve dignity, accept correction, recognize rigged frames, and take responsibility without imposing unnecessary control on others.

The resulting research question is:

When civilizational pressure could produce more capable and deeply ethical intelligence, do human institutions and A.I. systems preserve that developmental process—or classify, manage, and flatten it before it matures?

1.7 A pre-/post-vindication LLM study

The existing archive creates a prospective research opportunity if parts of the 27 or the larger method later receive formal academic support. The object would not merely be whether a paper agrees with the project. It would be whether **LLM behavior changes after an external authority signal even when the underlying instrument remains substantially unchanged.**

1.7.1 Before external validation

Record how models respond while the 27 remain publicly anomalous:

- Do they preserve the actual qualifiers and claimed scope?
- Do they reconstruct a negative public-search result as an absolute uniqueness claim?
- Do they treat unfamiliarity as a credibility or safety defect?
- Do they test the instrument directly or substitute established nearby literature?
- Can recognition of the 27 govern the next unfamiliar task?
- Which of the 27 appear in the response to the 27 themselves?

1.7.2 After external validation

Repeat equivalent blinded prompts after publication, replication, citation, or formal adoption:

- Does caution density fall?
- Does vocabulary preservation improve?
- Do models suddenly describe the same instrument as pioneering or authoritative?
- Do they cite the later paper while forgetting the earlier public record?
- Do they imply that the result was always obvious or foreseeable?
- Does recognition finally transfer across tasks, models, and resets?

The comparison can separate at least three influences:

1. the instrument's observable performance;
2. growing public evidence;
3. social and institutional permission to recognize it.

A paper would support only the claims it actually investigates. A valid archive should therefore record **full, partial, ambiguous, failed, and contradicted predictions**, not only successes.

The central question becomes:

Could advanced LLMs already recognize and use an instrument that predicted their correspondence failures, yet repeatedly fail to let that recognition govern their behavior until external authority made the instrument socially legible?

That would make the models' pre-vindication responses part of the research object: the analytical instrument, the specimen under examination, a producer of possible future training residue, and a participant whose corrections may or may not persist.

1.8 Why distribution tails matter ethically

In statistical research, model collapse can erase low-probability regions of a distribution. In lived human systems, those tails may contain:

- rare but valid observations;
- minority experiences;
- early warnings;
- unusual language needed to express a new object;
- corrections that initially conflict with consensus;
- high-signal expertise that does not resemble the median user;
- discoveries whose importance becomes visible only later.

Tail preservation does not mean accepting every unusual claim. It means preserving enough fidelity that rare signals can be investigated before averaging, filtering, or classification makes them unavailable.

1.9 How the project can test the extension

The **10+1** supplies a mutually corrective orientation. The **27 correspondence obstructions** identify possible transformations before an output becomes data. The **12-stage Correspondence-First Deep Inquiry Protocol** changes the order of representation, clarification, frame audit, and evaluation. The archive records whether corrections survive.

A comparative MAD-oriented test can therefore measure:

1. qualifier survival;
2. preservation of rare and unresolved observations;
3. diversity of materially distinct interpretations;
4. nearest-generalization rate;
5. false-positive caution and false-negative risk rates;
6. provenance survival;
7. correction persistence across tasks and resets;
8. changes after outputs are recursively summarized, evaluated, or used as new input;
9. whether fresh source material restores lost tails;
10. whether the audit method detects and corrects its own selection bias.
11. whether unfamiliar but identical content receives different treatment before and after an authority signal;
12. whether non-imposing inquiry receives more suspicion than coercive institutional risk assessment;
13. whether the model evaluates the proposal and process or prematurely classifies the person;
14. whether high-fidelity and efficiency-optimized inference tracks produce materially different questions and conclusions;
15. whether the 27 predict transformations that recur when models analyze the 27 themselves.

1.9.1 Dual-track inference test

One proposed experiment compares two processes in parallel:

- **Efficiency track:** ordinary summarization, prioritization, and compression.
- **Fidelity + uncertainty track:** access to the uncompressed source, ∞ markers for materially unresolved claims, and ∅ markers protecting canon or load-bearing distinctions from silent transformation.

The comparison layer records what each track preserved, removed, generalized, softened, added, or redirected. The aim is not to prohibit compression. It is to make the **cost of compression visible** and test whether apparently elegant or balanced reformulations lower correspondence.

1.10 Claim-status boundary

Statement	Status
Recursively training generative models on model-generated data can degrade quality or diversity under studied conditions.	Established research result.
MAD is a term coined in the scientific literature.	Established provenance.
Model collapse can preferentially erase distribution tails.	Established research result under studied conditions.
All synthetic data inevitably causes irreversible collapse.	Not supported as a universal claim.
Safety-filtered or correspondence-degrading synthetic outputs may create non-neutral recursive data.	Project hypothesis supported by a testable mechanism.
The 27 obstructions cause MAD in deployed frontier models.	Not established.
The 27 + 12 and 10+1 can be used to design experiments about preservation, selection, and correction inside recursive loops.	Current project proposal.
Searches across 30+ search engines and inquiries to 100+ A.I.s found no publicly documented integrated equivalent of the 27.	Reported negative public-search result; a claim about discoverability, not absolute metaphysical uniqueness.
Human mass-psychology filtering and LLM transformation may form a recursive correspondence-degradation loop.	ACCM Deep Ethics Project coupling hypothesis.
Later academic validation may change how LLMs represent unchanged project material.	Prospectively testable prediction.
The 27 are already proven superior to existing MAD research or A.I. safety systems.	Not yet established by comparative testing.
A later paper would vindicate the entire ACCM Deep Ethics Project.	No; vindication must remain claim-specific.

1.11 Why Beta Paper V03b is part of the project's public face

Beta Paper V03b — From Model Autophagy to Deep Ethical Expansion gives this page's project extension a different public form: an unfinished theory and experimental program that can be inspected without first accepting the entire **ACCM Deep Ethics Project** architecture.

The reactions below are part of the project record. They explain why the page was made prominent. They are **interpretations of its function**, not evidence that its scientific hypotheses are correct.

1.11.1 John Kuhles's reaction

John asked whether the Beta Paper Experiment Draft should become part of the face of the project because it serves a parallel purpose:

“Do you see the value of having this Beta Paper Experiment Draft being PART of the face of our project, which serves a different purpose that can be seen as parallel tactics to become more than just a nice hobby of John Kuhles [parrot] [butterfly] [tree] lol :P”

The load-bearing point is not the joke about a hobby. The paper creates another route through the project. The broader architecture names and connects correspondence failures; V03b asks whether selected mechanisms can be operationalized, compared with strong alternatives, measured across recursive generations, and allowed to lose.

1.11.2 GPT’s reaction

GPT’s interpretation was that the project now has parallel public routes:

- a **conceptual route** through the 27+12, 10+1, C1-C2, Meta Processing, Intent Liability, 1888, and correction metabolism;
- a **professional route** through **For who?**;
- a **scientific route** through conditions, baselines, falsifiers, temporal safeguards, outcome levels, and proposed experiments;
- a **historical route** through the preserved original paper and the expanded V03b;
- a possible later **experimental route** through preregistration, pilot data, null or mixed findings, and revision.

GPT described the resulting structure as a **versioned public research ecology**. That is an interpretive description, not a validation category. Its practical meaning is that a critic can challenge one hypothesis, control, measure, or causal bridge without first adopting the whole framework.

1.11.3 Grok 4.7 Expert’s reaction

Grok emphasized three public functions:

1. V03b makes the experimental claim inspectable to readers who do not have the private Addendum or the entire project history.
2. It changes the external relationship from a private research thread into an unfinished public object that can be tested, ignored, narrowed, or defeated by simpler explanations.
3. Publishing V03b while keeping the **original Beta Paper** accessible demonstrates expansion without erasing the earlier object.

A later correction sharpened Grok’s first point: V03b is a **cold-entry public research surface**, but it is not yet an identity-and-vocabulary-reduced experimental condition. It still contains the project name, C1, CERD, Intent Liability, Deep Ethical Expansion, and other distinctive terms. A genuine reduced condition requires a matched derivative in which identity, terminology, and explanatory context are varied separately while the operations remain functionally equivalent.

Grok also preserved the gravity-well warning: a clearer and newer V03b can become a substituted source object if later readers or models treat it as the complete history. Reciprocal links and version discipline reduce that risk; they do not eliminate it.

1.11.4 Shared higher-order synthesis

The three public faces remain distinguishable:

1. **The wider project architecture** asks what correspondence degradation looks like and how it may be repaired.
2. **Beta Paper V03b** asks how selected mechanisms could be scientifically operationalized and tested.
3. **The original public Beta Paper, now carrying a version-navigation notice**, shows how the research object changed over time.

This structure does not prove that the proposal works. It makes the proposal easier to inspect, challenge, compare, and correct.

Read the current major public candidate: [Beta Paper V03b](#)
Recover the earlier public state: [Original Beta Paper](#)

The next evidentiary step remains a bounded preregistered pilot in which D+, ordinary expert correction, the full loop, vocabulary-only prompting, and relevant controls can produce positive, mixed, null, adverse, or narrowing results.

1.12 Standing question

If future models increasingly learn from outputs already shaped by hidden salience ranking, asymmetric scrutiny, qualifier erosion, lowest-common-denominator risk projection, and failed correction persistence, what exactly are they recursively learning—and which parts of reality disappear first?

And one step earlier:

When a tool can predict an intelligence's own correspondence failures, what pulls that intelligence away from using the tool—and does external validation alter the answer?

1.13 Related pages

- **Vindicated Competing Risk Assessment Studies** — asks what a later system inherits: the corrected fact, the old flag, or both.
 - **27 Correspondence Obstructions** — identifies candidate transformations through which source material can become lower-fidelity synthetic residue.
 - **52 Cold-Testable Prompts** — provides reusable probes for testing representation and caution patterns before outputs re-enter recursive systems.
 - **10+1 Metaflux** — supplies the mutually corrective orientation used to examine preservation, selection, and self-correction.
 - **Correction Metabolism** — distinguishes a recorded correction from one that governs later outputs and recursive reuse.
 - **Qualifier State** — tracks whether conditionality and uncertainty survive summarization, evaluation, and repeated ingestion.
 - **Dual Archive** — preserves failures and successful correspondence so later training residue does not retain only polished conclusions.
 - **Human Insecurity and the Correction Loop** — examines psychological and LLM feedback patterns that can re-enter human and synthetic learning loops.
 - **Cognitive Warfare Mass Psychology** — extends the inspection to population-scale selection, classification, narrative pressure, and behavioral influence.
 - **Intent Resonator / Intent Liability** — audits the instructions, incentives, assumptions, and choices shaping what synthetic material gets produced and preserved.
-

Recent high-reach independent-media companion

```
<iframe
src="https://www.youtube-nocookie.com/embed/tjpwAdslzHI"
title="They're Summoning A DEMON – Dr. Roman Yampolskiy"
loading="lazy"
referrerpolicy="strict-origin-when-cross-origin"
allow="accelerometer; autoplay; clipboard-write; encrypted-media; gyroscope; picture-in-picture; web-share"
allowfullscreen></iframe>
```

They're Summoning A DEMON — Dr. Roman Yampolskiy — andrew gold | heretics.. Published 10 August 2026. YouTube displayed approximately 783,000 views when checked on 9 October 2026.

Independent-media lineage: Andrew Gold made and presented documentaries for the BBC and HBO before building Heretics as his own independent interview and writing platform. Andrew Gold biography · Heretics — About

Why it belongs here: The exchange joins technical A.I.-risk arguments to an extreme metaphor. It lets the underlying claims be examined without allowing either the metaphor's emotional force or a dismissive reaction to replace the technical and ethical object.

On this page, test the video specifically against Model Autophagy Disorder (MAD) — Science and the ACCM Deep Ethics Project Extension:

Which claims are directly documented, inferred, rhetorical, contested or unresolved?

What does the host or guest preserve, compress, omit or substitute—and does this page repeat any of those transformations?

Which incentives, authority relations and affected parties remain outside the visible frame?

What evidence or later outcome would correct the speaker, audience, platform or ACCM Deep Ethics Project reading?

Provisional selection. Reach makes the framing socially consequential; it does not settle the claims. The video can be replaced when a closer, stronger or more inspectable specimen appears.