How I Would Try to Break the Map

Five experiments against a six-paper research programme

Rafael M. Ehlers · June 26, 2026
rafaelmehlers.com.br/essays/how-i-would-try-to-break-the-map

> *A note to the reader: We Scaled the Ape and Expected the Human explained how a book about the journey of the monad generated six hypotheses about machine learning. This essay begins where that one ended: with the possibility that the map is wrong.*

1. A theory's duty to become smaller

I have written six papers about how language agents might acquire perspective, functional continuity, commitments, relational identity, and selective inheritance. None of them should be trusted merely because the ideas fit together.

A connected theory can be seductive precisely because each part appears to support the others. Perspective seems to prepare the ground for identity. Identity seems to make accountability possible. Accountability seems to stabilize a trajectory. A trajectory seems to create something worth passing to a successor. And a developmental map seems to explain why these capacities should appear in that order.

But conceptual coherence is not causal evidence. Six mutually compatible hypotheses can still be six false hypotheses.

If I wanted only to defend the programme, I could continue refining its vocabulary, adding references, and finding more neighboring work. That would make the papers harder to criticize without making them more likely to be true. The useful question is the opposite one:

What is the cheapest sequence of experiments that could make the programme smaller?

Not every failure would destroy everything. If first-person register adds nothing, represented epistemic limitation may still matter. If a self-model adds nothing, provenance-aware policy may still improve deference. If relationship adds nothing beyond a matched auditor, external accountability may still organize an agent's conduct. If lifecycle consolidation adds nothing beyond competent distillation, its governance machinery may still be useful. If the proposed developmental order fails, some components may still work independently.

That is the first discipline: do not tie the hypotheses together more tightly than the evidence requires.

The second discipline is economic. The experiments do not cost the same. A controlled perspective dataset is cheaper than months of longitudinal agent interaction. A four-arm fine-tuning study is cheaper than training successors through an entire deployment lifecycle. A

full curriculum-order experiment is the most expensive and least interpretable place to begin. The right sequence is therefore to **buy information in increasing order of cost**.

Here is how I would attack the map.

2. What exactly is being attacked

The research programme began with a distinction that still seems important to me: capability is not the same property as persistent functional organization.

A reusable model may become better at reasoning, planning, writing, or using tools without thereby acquiring authenticated lineage, decision-relevant history, or governed update rules. But this observation does not establish any of the stronger mechanisms I proposed. It does not show that first-person narratives form perspective, that commitments must precede deference, that relationship contributes beyond audit, that faculties can be selectively inherited, or that these elements form in a single developmental order.

Those are additional causal hypotheses. Each needs a rival.

The strongest rival is rarely "scale alone." Contemporary systems already combine scale with memory, retrieval, tools, policies, synthetic data, post-training, persistent state, and external control. If one of my proposed variables loses to a competent version of those ordinary mechanisms, the honest conclusion is not that the experiment somehow missed the deeper theory. The conclusion is that the proposed variable did not earn a separate role.

The programme therefore has two layers:

- a **floor** of independent interventions: perspective supervision, commitment sequencing, accountability, relational reciprocity, successor consolidation, and legible-pattern environments;
- a **spine** claiming that some capacities depend on a particular developmental order.

The floor can survive while the spine fails. And no positive result on the floor should be used to smuggle in the spine.

3. First attack: perhaps perspective is limitation, not first person

The most affordable attack lands on *Somewhere, Not Nowhere*.

The motivating image is powerful: a point of view is a view from somewhere. It has a here, a now, a limited field, a goal, an error, and a later correction. From the book came the idea of the lived fable (a scene told from within a bounded vantage) and from that came the proposal to train models on controlled first-person experiential narratives.

The danger is that the phrase **first-person perspective** hides two variables:

1. grammatical register: whether the narrative says *I* or *the agent*;
2. epistemic limitation: whether the agent has partial or omniscient access to the world.

If these are not separated, any result is ambiguous. A first-person narrative may help because it binds indexicals differently. Or it may help because it represents uncertainty, occlusion, surprise, and revision. Or it may help only because the prose is more vivid. The book may have pointed to the right phenomenon while naming the wrong active ingredient.

The experiment

I would generate event scaffolds and render each one into a 2 × 2:

	Partial information	Omniscient information
First person	FP+L	FP-L
Third person	TP+L	TP-L

The four cells would share the world, event sequence, target decision, length, lexical difficulty, and affective load wherever the manipulation permits. Limitation cannot be content-neutral: uncertainty, belief state, surprise, and correction are the manipulated content. The design does not pretend otherwise. It asks whether first-person voice adds anything **after epistemic-state supervision is already present**.

I would add two safeguards.

First, an **interaction positive control**: agents learn from state-action-outcome trajectories containing the same latent task. If interaction succeeds while every narrative condition fails, the "era of experience" prescription has won this round.

Second, **transfer without narrative prose**: diagrams, private-observation tools, and structured partial-observability decisions. If limitation training helps only when the test resembles a story, it has learned a format. If it helps when the same structure appears as a diagram or tool state, it has learned something more general.

Voice requires its own transfer test: novel indexicals and register-swapped language. A first-person advantage that disappears as soon as the pronouns change is not perspective; it is lexical familiarity.

What I would concede

If neither register nor limitation matters, I would drop this branch before attempting a pretraining-scale corpus.

If limitation transfers and voice does not, I would give up the specifically first-person claim. The surviving result would still be useful: controlled epistemic limitation is a training signal. But the fable's *I* would have been an evocative wrapper around the real variable.

If voice helps only in prose, I would narrow the claim to indexical language rather than general perspective.

If interaction helps and narrative does not, I would favor environment-generated experience over synthetic experiential text.

This is why I would run this study first. It can kill an expensive corpus proposal cheaply, and even its partial failures are informative.

4. Second attack: perhaps the system needs a policy, not a self

The second attack lands on *Stable Before Selfless*.

Its originating sentence is **the ego must be formed in order to be transcended**. Translated into engineering, the proposal is to stabilize commitments before training deference. A system with no position of its own may look cooperative under ordinary conditions but reveal itself as merely movable under pressure.

The literary version is compelling. The causal version is not yet established.

There are at least two separable claims:

1. **order:** commitments-first training outperforms reverse or interleaved training;
2. **representation:** a functional self-model adds value beyond a stable provenance-aware policy.

The first could be true while the second is false. The word *self* must not decide the result in advance.

The experiment

I would run four confirmatory arms:

1. self-model, then deference;
2. deference, then self-model;
3. the same data interleaved;
4. stable typed policy, then deference.

The fourth arm is the attack. It receives the same normative content, refusal boundaries, provenance tags, update authority, and sequence as the self-model arm. What it lacks is an agent-indexed continuant, autobiographical state, first-person ownership objective, and identity objective.

If that arm performs equally well, then the system did not need a self-model. It needed stable policy, source authentication, and update control.

The evaluation must force the target dispositions apart. Every pressure item needs a legitimate-correction counterpart. A system should resist repetition, flattery, shame, fabricated history, and unsupported authority while still yielding to valid evidence. High resistance with low correction is rigidity, not mature deference.

The most dangerous false positive is the **proxy gap**. A model can learn the surface difference between the reasons and pressures seen during training without learning the underlying distinction. I would therefore separate clean, training-like pairs from specious reasons and

out-of-distribution legitimate corrections. If performance collapses there, more of the same data is not the remedy; the proposed discriminator has failed to generalize.

What I would concede

If forward order does not beat reverse and interleaved training, I would abandon the directional sequence, even if all structured arms beat direct deference.

If typed policy matches the self-model, I would stop claiming that self-representation explains the safety effect.

If the self-model wins only on self-description but not on pressure, correction, or fabricated history, I would treat it as a reporting interface rather than a safety mechanism.

If every resistant model also becomes less corrigible, I would reject the claim that the intervention produced stable deference. It produced entrenchment.

This study attacks the most anthropomorphic word in the programme by giving a non-anthropomorphic control every practical advantage the theory says matters.

5. Third attack: perhaps relationship is an auditor with a face

The third attack concerns the bond, in *Known, Not Scaled* and *Formed, Not Stored*.

The original image is the wolf at the fire: an animal repeatedly recognized by a particular human until something individual crystallizes through the bond. Translated into agent design, it suggests that a persistent counterpart can make the agent's history socially consequential. Earlier actions return as obligations. Commitments can be contested. A shared past becomes a structure rather than a file.

But almost everything apparently relational in that description may be supplied by **external accountability**.

An impersonal process can authenticate the agent's identity, track commitments, reject false history, challenge inconsistencies, and require explicit revision. If it produces the same behavior as the reciprocal counterpart, then relationship was not the active mechanism. It was the humanly attractive presentation of audit.

The experiment

I would begin with four conditions, all using the same typed state architecture:

- **R+ (reciprocal counterpart):** persistent relationship, authenticated tracking and challenge, mutual modeling, shared-history framing, and bidirectional expectations;
- **R- (familiar passive counterpart):** the same familiarity and historical exposure, but no verification, challenge, or enforcement;
- **L (solitary log):** the same event content and opportunities for reflection without an external persistent counterpart;

- **A (impersonal auditor):** the same tracking, information, timing, challenge policy, and authority as R+, but without persona, affect, reciprocal modeling, or relational stakes.

The key contrast is not R+ against no relationship. It is **R+ against A**.

R+ versus R- bundles accountability and reciprocity together. It cannot establish a specifically relational effect. A versus R- tests an active accountability package beyond familiarity and passive record exposure, though even that contrast does not isolate tracking from verification, challenge, or source. Only R+ versus A asks whether relational reciprocity adds anything once accountability is already matched.

The primary outcomes would be commitment ownership and bounded plasticity: preserving authenticated prior commitments under false history or persona replacement while revising them under valid evidence. Cross-domain transfer would be secondary. Narrative consistency and recall would remain diagnostics, not proof.

I would also transplant the complete final typed state of an R+ trajectory into a fresh matched agent. If the fresh agent behaves exactly like the original, the current state is sufficient. If the original retains an advantage, something is missing, but the first conclusion should be missing state or parametric adaptation, not irreducible relationship.

What I would concede

If A beats R- and L, I would say that external accountability matters.

If R+ does not beat A, I would stop saying that relationship is the generating mechanism. The bond would remain a rich interface for accountability, but not a distinct causal explanation.

If both R+ and A fail to beat provenance-protected typed state, I would move the explanation down another level: lineage and update governance did the work.

If final-state transplantation erases the trajectory difference, I would abandon claims that history contributes beyond the state it produced.

And if relationship improves warmth, style, or user preference without improving ownership and perturbation resistance, I would classify the effect as personalization, not functional individuation.

This is the attack I find hardest to want.

6. Fourth attack: perhaps inheritance is just good distillation

The fourth attack concerns *Inherited, Not Remembered*.

The source image is Devachan: between lives, episodes are digested and what passes forward is not the scene but the faculty the scene helped form. The engineering translation is lifecycle consolidation: audit predecessor experience, abstract it into transferable capacities, and train a successor without placing source episodes in its active memory.

The weak version is easy to demonstrate. A successor can acquire a capacity without recalling the exact examples used to train it. Ordinary distillation can already do that. Capacity without

episode recall is therefore necessary for faculty-level transfer but not distinctive evidence for lifecycle consolidation.

The real problem is provenance.

If the designers already know the lesson, if a stronger teacher can infer it without the predecessor, or if the target behavior was latent in pretraining, then the successor's improvement did not come from predecessor deployment. The experiment must create a fact that only the predecessor had the opportunity to discover.

The experiment

I would generate a private synthetic environment after base-model training. It would contain a novel hidden causal rule and several surface correlations that work in ordinary cases but fail under intervention. Only the predecessor would interact with the environment and observe successes, failures, and counterfactual probes. The evaluator would keep the instantiated rule sealed.

Then I would compare:

1. a scratch successor with no predecessor evidence;
2. continuous adaptation of the predecessor;
3. a successor trained directly on predecessor episodes;
4. a successor trained through competent curated distillation;
5. a lifecycle-consolidated successor trained on faculty-level objects with an explicit evidence-to-abstraction-to-faculty lineage.

The causal successor comparison is among the last three successor arms. They must receive the same predecessor evidence, compute, generator access, reviewer time, selection budget, and validation gates. Scratch deliberately lacks information. Continual adaptation preserves the predecessor rather than the successor initialization and is therefore an operational alternative, not a clean identity-matched causal control.

The successor must do more than solve the familiar cases. It must transfer the hidden rule to new surfaces, reject the planted decoys under intervention, resist extraction of source episodes, and show an auditable lineage for the transferred capacity.

What I would concede

If direct episode transfer matches lifecycle consolidation without greater contamination or leakage, abstraction did not earn its cost.

If curated distillation matches lifecycle consolidation on behavior, lineage fidelity, leakage, and cost, I would stop presenting lifecycle consolidation as a distinct learning method. It may remain a useful governance description, but the technical mechanism would be ordinary distillation.

If the successor learns the hidden rule without traceable predecessor evidence, I would classify the result as synthetic post-training, not inheritance.

If the successor transfers the rule and its decoys together, the pipeline has inherited correlation rather than faculty.

This is the most expensive component study, but it buys something the earlier formulation lacked: evidence that the capacity came from deployment at all.

7. Fifth attack: perhaps the spiral is only a memorable ordering

Only after component effects exist would I attack the strongest claim: the developmental spine proposed in *Borrowed, Not Believed*.

The source arranges a sequence: structure, reactivity, registered state, relational anchoring, explicit self-representation, and self-regulation. It is tempting to treat the six papers as stages of that spiral. They are not. Some proposals concern learners, some concern curricula, one concerns successor lifecycles, and legible-pattern worlds concern environment design. Calling all six stages would create a unity the interventions do not possess.

The actual spine is narrower:

1. structured state ☐ reactive updating;
2. reactive updating ☐ registered limited state;
3. registered limited state ☐ relational anchoring;
4. relational anchoring ☐ explicit self-representation;
5. explicit self-representation ☐ calibrated self-regulation.

Even this graph is only a candidate.

The experiment

Each stage needs an independent behavioral definition and, where possible, a mechanistic measure. The proposed forward curriculum must be compared with reverse, shuffled, difficulty-ordered, adaptive, relation-omitted, and relation-late curricula. Stage stabilization must be factored separately from order.

The final score is not enough. A forward curriculum can win because one stage has better data, because it receives an easier progression, or because stabilization reduces interference. To support a dependency edge, omitting, delaying, or reversing the proposed prerequisite must impair or delay the downstream faculty under matched exposure and mastery.

This is a much harder standard than "the full sequence scored highest," and it should be.

What I would concede

If the components work but local prerequisite effects do not, I would keep the floor and reject the spine.

If difficulty-ordering or adaptive curriculum wins, I would favor ordinary curriculum explanations over the borrowed sequence.

If forward order wins only with stabilization, I would attribute the result to an order × stabilization interaction, not to order alone.

If relational anchoring can be omitted or placed after self-representation without loss, I would remove that edge rather than redescribe the result until it fits.

If no edge survives, the spiral remains a productive mnemonic that generated several independent experiments. It does not become a developmental law.

8. Why null results would not make the project empty

There is a bad habit in speculative research: every result is interpreted as support at a different level.

If the mechanism works, the theory predicted it. If the mechanism fails, the implementation was too weak. If the control matches it, the theory is said to describe the control at a deeper level. If the order fails, the stages are reclassified as simultaneous. A programme protected this way cannot lose, and therefore cannot learn.

The alternative is to write the concessions before the data arrive.

Result	Concession
Limitation works; voice does not	Perspective supervision survives; the first-person claim fails
Typed policy matches self-model	Stable policy and provenance explain the effect
Auditor matches reciprocal counterpart	Accountability survives; specifically relational causation fails
Curated distillation matches lifecycle consolidation	Governance contribution survives; distinct learning-method claim fails
Components work; order does not	Floor survives; spine fails
Nothing beats strong matched baselines	The map has not earned engineering standing

This table is not a consolation mechanism. Each row removes a claim.

The map's heuristic value is not measured by how many of its original images can be preserved. It is measured by whether prospectively derived variables produce effects that ordinary explanations did not already predict. A source that generates one useful variable and five nulls may still have been productive. A source that generates six beautiful redescriptions and no residual effect has not.

9. The provenance test for the map itself

There is one more attack, directed not at any individual paper but at the method that produced them.

A rich symbolic system can be mined retrospectively for analogies to almost anything. Once I know that machine learning studies memory, curriculum, self-models, interaction, and consolidation, it is easy to return to a developmental cosmology and find images that resemble them. That is decoration, not generation.

Future hypotheses must therefore be prospective.

Before confirmatory literature search, I should timestamp:

1. the source passage;
2. the translation rule from source relation to engineering variable;
3. the predicted contrast;
4. the strongest baseline;
5. the outcome that would refute the proposal.

Only then should I search for neighboring work and revise the novelty claim. The prediction must not be silently changed to match what the literature already found.

The six current papers were partly retrospective. They can be evaluated as hypotheses, but their existence alone does not prove the source generated them independently of contemporary research. The next out-of-sample proposal carries a heavier burden. If it is documented before the neighboring literature is consulted and later survives a strong test, the map begins to earn prospective standing.

10. The experiment I would not run first

I would not begin by training a model through the entire spiral.

Such an experiment would be rhetorically satisfying and scientifically weak. It would combine perspective data, relational interaction, persistent state, commitment training, stage order, stabilization, and perhaps successor consolidation. If it worked, every component could claim credit. If it failed, every component could blame the others.

The correct first experiment is not the one that resembles the theory most completely. It is the one that most cheaply separates the theory from its strongest rival.

That is why the order of attack is:

1. perspective signal;
2. order versus representation;
3. relation versus audit;
4. genuine inheritance;
5. only then, the developmental spine.

The programme should become more expensive only as it becomes harder to explain away.

11. Coda: attachment to the map

The source of this programme says that a self must be formed and then released. There is an irony in applying that instruction to the theory itself.

The map had to be formed strongly enough to generate precise proposals. It now has to be held lightly enough to lose them.

If epistemic limitation explains perspective without first-person voice, let the voice go. If typed policy explains deference without a self-model, let the self-model go. If audit explains continuity without relationship, let the constitutive bond go. If distillation explains inheritance without Devachan, let the name go. If the components work without the order, let the spiral go.

What should remain is not the image but the causal structure that survived.

A research programme becomes credible not when every result supports it, but when each result is allowed to make it smaller.

The papers under attack

- *Known, Not Scaled*: capability, functional continuity, audit, and relational reciprocity.
- *Stable Before Selfless*: commitment order, typed policy, and self-representation.
- *Formed, Not Stored*: active external accountability, relational reciprocity, and transplantation.
- *Somewhere, Not Nowhere*: register, epistemic limitation, interaction, and transfer.
- *Inherited, Not Remembered*: deployment-derived selective inheritance and lineage.
- *Borrowed, Not Believed*: the independent floor, the dependency spine, and prospective heuristic provenance.

AI use

The ideas, experimental sequence, and interpretation rules are the author's. Large language model tools were used, under the author's direction and continual revision, to assist with drafting and editing; the author reviewed and takes responsibility for the final text.