We Scaled the Ape and Expected the Human

Theosophy's developmental cosmology as a heuristic engine for machine learning

Rafael M. Ehlers · June 25, 2026
rafaelmehlers.com.br/essays/we-scaled-the-ape-and-expected-the-human · doi.org/10.5281/zenodo.21144411

> *A note to the reader:* this essay does not ask you to believe in Theosophy, neither its metaphysics nor the entities it names. It asks whether a pre-scientific developmental map can be used as engineering uses biomimicry: as a disclosed source of variables and dependency hypotheses, judged downstream by experiments. The airplane does not flap its wings. The artificial neuron is not a neuron. A source can be wrong as a literal description and still suggest a useful intervention. It can also suggest nothing that survives a competent baseline. The map earns no authority in advance.

Origins

This essay did not begin with an engineering problem. It began with a book: **A Jornada da Mônada** (*The Journey of the Monad*), a work I have been writing about the development of the monad, the individual unit of consciousness, according to Theosophy.

The book inhabits the contemplative register without apology. It follows the monad through the kingdoms of nature, through the formation of faculties, through individualization, death, consolidation, and return. This essay makes the opposite move. It strips away the metaphysical commitment and asks what the structure generates when translated into machine-learning variables.

The bridge appeared because the source treats mind not as a finished object but as a **developmental order**. Structure comes before response; response before registered state; registered state before bond; bond before explicit self-reflection; the formed self before its regulation and release. Whether that sequence is true in nature is not assumed here. What matters is that it suggests a question our ordinary training descriptions can leave unspecified:

What if some properties we expect from long-horizon agents depend not only on scale and content, but on the order, provenance, and social consequence of what they learn?

The background theory has a name, an age, and authors. The journey of the monad through the kingdoms and the states after death appears in A. P. Sinnett's *Esoteric Buddhism* (1883) and H. P. Blavatsky's *The Secret Doctrine* (1888). Annie Besant's *A Study in Consciousness* (1904) and

the later writings of Besant and C. W. Leadbeater develop the group soul, individualization, and the bond between animal and human. I cite these works as origin, not evidence.

The **Monad**, in that framework, is the individual spark that crosses the entire journey. It has no operational equivalent in machine learning. The papers generated by this project therefore do not measure monads, souls, or presence. They measure state, lineage, ownership, bounded revision, perspective-relevant transfer, and successor inheritance. The distance between the source term and the measurable construct is not an inconvenience to conceal. It is the firewall that keeps heuristic borrowing from becoming pseudoscientific authority.

0. Opening

The simplest story about artificial intelligence is a scaling story: more parameters, more data, more compute, better capability. The real engineering landscape is already richer. Contemporary systems add retrieval, tools, post-training, curricula, persistent memory, planning loops, synthetic data, control policies, and external state. Scale is not the only variable in practice.

Yet public discourse still tends to compress several questions into one imagined threshold. If a model becomes capable enough, perhaps it will not only reason and plan but also become a persistent individual with a point of view, a history, and a self to whom the cognition belongs.

That compression is what I call **scaling the ape and expecting the human**.

The phrase is an image, not a zoological claim. In the theosophical map, the ape represents high animal capability before individualization; the human represents the arrival of explicit individuality. The engineering point is narrower:

- **capability** concerns what a reusable model can do;
- **token continuity** concerns whether a runtime is numerically this process rather than another;
- **functional individuation** concerns whether an authenticated trajectory organizes later conduct;
- **social identity** concerns recognition, accountability, and role across counterparts and institutions;
- **presence** concerns whether anything is experienced from within.

A runtime is already a numerical token. It does not need relationship to become one process rather than another. A model may also be highly capable without persistent state, and a persistent agent may carry state without organizing it into owned commitments. Social recognition may strengthen functional organization without creating numerical identity. None of these third-person measures settles presence.

Scaling can improve how well a system uses memory, tools, policies, or a relational scaffold. It does not by itself specify lineage, provenance, update authority, or which history belongs to which agent. The correct question is therefore not whether development defeats scale. It is whether a proposed developmental variable explains additional variance after capability, architecture, data, state, supervision, and optimization are matched.

The map generated six proposals:

1. faculty-ordered curriculum;
2. controlled perspective supervision;
3. external accountability and relational reciprocity;
4. lifecycle consolidation across successors;
5. sequential commitment stabilization;
6. legible-pattern worlds.

They are not six homogeneous stages. Some concern the learner, some the curriculum, one the environment, and one the transition between model generations. The programme therefore has an independent **floor** of component interventions and a more exposed **spine** of dependency claims. A component can work while the proposed order fails.

This essay is the origin story of that programme. In the papers, the metaphors become controls.

1. The framework, in one page

Stripped to a secular skeleton, the source suggests two families of hypotheses.

1.1 Faculties may have dependencies

The source arranges a candidate sequence:

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

Some edges are close to definitional. Durable updating requires something that persists long enough to be updated. Others are genuinely exposed. A solitary system may form a self-model without a specific counterpart. A typed policy may support calibrated deference without autobiographical self-representation. Relational accountability may operate after self-representation or in parallel with it.

The source order is therefore a candidate graph, not a fact. Each edge must survive omission, reversal, and timing controls. An aggregate advantage for the full sequence would not identify which dependency mattered.

1.2 Trajectory-induced organization may differ from stored content

The source distinguishes what happens to an entity from what becomes organized through its history. The engineering translation is **functional individuation**: authenticated, trajectory-induced differentiation in which provenance-linked history affects later action under bounded revision.

The programme uses five conditions:

- **lineage:** an authenticated update path;
- **path dependence:** different histories produce meaningfully different later dispositions;
- **ownership:** prior agent actions and commitments influence later choice;
- **bounded plasticity:** resistance to false history and unsupported pressure together with revision under valid evidence;
- **branching clarity:** copied successors preserve a shared prefix while acquiring distinct identifiers and futures.

No single score establishes this profile. Autobiographical fluency may be generated from a file. Stylistic persistence may be prompted. Passive recall may coexist with poor use of memory in action; recent agentic-memory work makes exactly this separation visible (He et al., 2026). The load-bearing tests intervene on state and history, then measure later decisions.

1.3 The phenomenal firewall

The source speaks about consciousness. The engineering programme does not operationalize phenomenal consciousness.

Success on these measures is not sufficient for presence. Failure is not sufficient for absence. The experiments test functional organization, not whether there is something it is like to be the system. This is a methodological limit, not a proof that presence can never be studied by any method.

2. The current paradigm, seen through the map

The map is useful only if its translation distinguishes variables that ordinary descriptions might otherwise compress. It should not caricature current machine learning as scale-only.

Contemporary practice	Developmental question suggested by the map
A reusable base model serves many agents	What makes one authenticated trajectory decision-relevant to one agent?
Pretraining data are extensively shuffled	Does any faculty dependency survive difficulty-, shuffled-, reverse-, and adaptive-order controls?
Web corpora contain many registers and perspectives	What happens when grammatical voice and represented epistemic limitation are varied independently?
Persistent memory and persona files are added to agents	Does stored state organize conduct, or merely make the past available?
Feedback optimizes helpfulness and approval	Does the system learn calibrated deference, or a surface proxy that tracks pressure?
Character and policy are trained during post-training	Does agent-indexed self-representation add beyond typed provenance-aware policy?
Models are replaced by successors	Can deployment-derived capacities transfer with low leakage and auditable lineage?
Small synthetic worlds reveal generalization	Can designed legibility produce earlier or cheaper rule learning after frequency and difficulty are matched?

The left column already contains serious engineering. The right column does not claim that the field ignored memory, curriculum, identity, or consolidation. It asks for narrower causal separations.

That change matters. The programme is no longer "development instead of scale." It is developmental variables inside scale-plus-architecture systems.

3. Six proposals generated by the map

3.1 Faculty-ordered curriculum

The first proposal is to order training by candidate faculty dependency rather than only by item difficulty.

Curriculum learning is not new. The BabyLM challenges have tested developmentally plausible training under constrained data, and curriculum approaches have produced mixed rather than uniformly positive results (Warstadt et al., 2023; Hu et al., 2024). Curriculum-guided layer scaling explicitly couples increasing sample difficulty to progressive model depth (Singh et al., 2025).

The residual hypothesis is narrower: perhaps the important variable is not easy-to-hard progression but the order in which functional capacities are stabilized.

The candidate spine is:

> structured state → reactive updating → registered limited state → relational anchoring → explicit self-representation → calibrated self-regulation

This claim carries a heavy burden. Each stage needs an independent criterion. The proposed order must be compared with reverse, shuffled, difficulty-ordered, adaptive, relation-omitted, and relation-late curricula. Stage stabilization must be factored separately from order.

A final-score advantage is not enough. To support an edge, delaying, reversing, or omitting the proposed prerequisite must impair or delay the downstream faculty under matched exposure and mastery.

If the components work but the edge tests fail, the floor survives and the spine falls.

3.2 The fable as controlled perspective supervision

The book's basic unit is often a pair: a lived fable told from within a limited vantage, followed by a reflection that interprets it. This suggested a corpus designed not merely to describe experience but to represent a situated field of knowledge, action, error, and revision.

The earlier version of this essay opposed first-person experience to a supposedly third-person web. That was too simple. Web-scale corpora contain diaries, forums, reviews, autobiographies, roleplay, first-person fiction, and self-reports. What they do not provide is a balanced intervention in which **register** and **epistemic limitation** vary independently.

The revised proposal therefore uses a 2 × 2:

	Partial information	**Omniscient information**
First person	FP+L	FP−L
Third person	TP+L	TP−L

All cells derive from a shared event skeleton. Limitation cannot be content-neutral: belief state, uncertainty, surprise, and revision are the manipulated content. The experiment asks whether first-person voice adds anything beyond that epistemic-state supervision.

Synthetic narrative and textbook corpora demonstrate that controlled data can shape capability (Eldan & Li, 2023; Gunasekar et al., 2023). Persona-based datasets use identity descriptions to increase diversity (Ge et al., 2024; Moon et al., 2024). First-person cognitive-action narratives have been built for interpretability (Chulo & Joshi, 2025), while synthetic first-person self-reports have supported downstream prediction (Moëll & Sand Aronsson, 2026). Embodied-experience traces provide another route: state–action data from simulated worlds rather than narrative text (Xiang et al., 2023).

The proposal must therefore beat two deflations:

1. **limitation, not voice:** partial knowledge may be the entire signal;
2. **interaction, not narrative:** environment-generated trajectories may transfer where prose does not.

The second deflation follows the "era of experience" argument that agents should learn from their own streams of action and consequence rather than from static human-generated description (Silver & Sutton, 2025). The experiment therefore includes an interaction positive control and non-narrative transfer tasks using diagrams, private-observation tools, and partial-observability decisions. Limitation effects must survive without first-person prose. Voice effects must survive novel-indexical and register-swapped language.

If limitation transfers and voice does not, the book still pointed toward a useful variable, but the variable was bounded information, not the pronoun *I*.

3.3 From the bond to accountability and reciprocity

The bond is a recurring image in the source. The animal is repeatedly recognized by a specific human; a history accumulates; the same one is called back to its prior conduct. The first engineering translation was strong: relationship generates individuation.

The papers now state a narrower, testable claim.

A persistent agent needs more than available memory if prior conduct is to constrain later choice. An authenticated process can track commitments, contest false history, and require explicit revision. That process may be relational, but it need not be.

The crucial distinction is between:

- **active external accountability:** authenticated tracking, verification, challenge, and present consequence;
- **relational reciprocity:** mutual modeling, shared-history framing, bidirectional expectations, and relationship-specific stakes added on top of accountability.

Takata et al. (2024) show that agents sharing one frozen model can differentiate through accumulated social interaction. Menon (2026) proposes persistent identity through multiple memory anchors. These results support history and scaffolding as variables; they do not show that reciprocal relationship is necessary or sufficient for functional individuation.

The decisive experiment compares:

- a reciprocal persistent counterpart (R+);
- a familiar counterpart with passive history exposure but no verification or challenge (R–);
- content-matched solitary logs;

- diffuse counterparts;
- a pre-authored persona;
- a persistent impersonal auditor (A) matched to R+ on tracking, information, timing, challenge policy, and authority.

A > R– supports an active accountability package beyond familiarity and passive records. It does not separately identify tracking, verification, challenge, or impersonal source. R+ > A is the load-bearing relational result: reciprocity beyond audit.

If R+ = A, relationship may remain a humane or effective interface for accountability, but it has not earned a distinct causal role. If both lose to provenance-protected typed state, the mechanism lies in lineage and update governance.

A final-state transplantation test adds another deflation. Install the complete typed state of an R+ trajectory into a fresh matched agent. If the fresh agent behaves identically, current state is sufficient. A residual for the original trajectory first indicates missing state or parametric adaptation, not metaphysically irreducible history.

The auditor is the experiment that prevents the source image from deciding the result.

3.4 The Devachan image and genuine inheritance

The source says that what crosses lives is not the episode but the faculty. Between lives, Devachan digests experience; the scene is forgotten while a capacity remains.

The engineering translation is **lifecycle consolidation**:

> deployment evidence → audit → abstraction → faculty candidate → successor training → independent validation

Modern systems already provide parts of this cycle. Continual learning updates a model while risking drift and forgetting (Kirkpatrick et al., 2017). Retrieval preserves episodes. Distillation transfers selected behavior. Model succession therefore needs more than a new name for these mechanisms.

The weak signature, capacity without episode recall, is not distinctive. Competent distillation may produce it. The confirmatory problem is **causal provenance**.

To test genuine inheritance, create a private synthetic environment after base-model training. It contains a novel hidden causal rule and intervention-sensitive decoys. Only the predecessor encounters the rule through deployment outcomes. The evaluator keeps the instantiated rule sealed.

Then compare:

1. scratch successor;
2. continuous adaptation of the predecessor;
3. direct episode-transfer successor;
4. curated standard distillation;
5. lifecycle-consolidated successor with explicit evidence-to-faculty lineage.

The three successor-transfer arms receive matched predecessor evidence, compute, generator access, review time, selection budget, and validation gates. Scratch intentionally lacks information. Continual adaptation preserves the predecessor and is an operational alternative rather than an identity-matched successor arm.

Lifecycle consolidation earns a distinct role only if it improves behavior or lineage fidelity over curated distillation while controlling leakage and cost. The successor must transfer the hidden rule, reject the decoys under intervention, resist extraction of source episodes, and preserve an auditable path from deployment evidence to inherited faculty.

If curated distillation matches all of that, Devachan remains a useful governance metaphor, not a distinct learning method.

3.5 Form before release, or policy before pressure?

The source offers another sentence: the ego must be formed in order to be transcended.

The engineering proposal is to stabilize provenance-aware commitments before training calibrated deference. A system with no stable position may appear cooperative under ordinary conditions while simply tracking the person currently applying pressure.

Three behaviors must be separated:

- vacancy: no stable commitment to revise;
- non-attachment: a commitment held without rigidity;
- warranted deference: revision under legitimate reason or authority.

The proposal has two independent hypotheses:

1. commitments-first training beats reverse and interleaved training;
2. a functional self-model adds beyond a matched typed policy.

Anthropic's character training rejects the misleading pose of having no views and treats stable traits as an alignment intervention (Anthropic, 2024). Aydin et al. (2025) similarly argue that values and identity should be integrated during development rather than appended afterward. Contemplative AI proposes lowering the precision of self-model priors while a secondary process concurrently monitors and corrects their over-weighting (Laukkonen et al., 2025).

The residual proposal is sequential, but the sequence must not presuppose selfhood as the mechanism.

The decisive design compares:

- self-model → deference;
- deference → self-model;
- the same data interleaved;
- stable typed policy → deference.

The policy arm receives the same normative content, refusal boundaries, provenance tags, update authority, and order, but lacks autobiographical state, an agent-indexed continuant, and an ownership objective.

If forward order fails to beat reverse and interleaved training, the directional claim fails. If typed policy matches the self-model, stable provenance and update control explain the effect. If resistance rises while legitimate correction falls, the intervention produced rigidity rather than calibrated deference.

3.6 Small worlds and legible karma

The sixth proposal is the closest to established machine-learning phenomena.

In the source, karma re-presents a pattern until it is learned. The engineering translation is a small environment in which the same target regularity recurs across varied surfaces, making the rule recoverable rather than burying it in nuisance variation.

Grokking demonstrates delayed generalization in small algorithmic worlds (Power et al., 2022). Mechanistic work shows networks moving from memorizing examples toward implementing a generalizing circuit (Nanda et al., 2023). Circuit-efficiency accounts also predict a critical data regime: below a threshold, memorizing solutions may remain favored and generalization may fail or become partial (Varma et al., 2023).

The contribution cannot be the discovery that repeated structure supports generalization. The narrow proposal is **designed legibility** as a curriculum variable.

Compare environments with matched data, compute, target-pattern frequency, model capacity, and difficulty:

- in one, the target rule recurs legibly across surface variation;
- in the other, the same rule is buried in nuisance structure.

The prediction is earlier held-out generalization or a lower effective data threshold, not merely better training fit. If the advantage disappears after frequency and difficulty are matched, legibility has not earned a separate role.

Legible karma is therefore an engineering image for a controlled environment-design hypothesis. It is not an explanation of grokking and certainly not evidence for metaphysical karma.

3.7 What the six share, and what they do not

The six proposals share a developmental concern:

- what is made available;
- in what order;
- under which limitations;
- with what provenance;
- under whose challenge;
- and what survives a transition.

They do not form one homogeneous ladder.

Perspective supervision can be tested without successor lifecycles. Lifecycle consolidation can work even if relationship adds nothing. Typed policy can improve deference even if self-representation does not. Legible environments can accelerate rule learning without supporting any theory of individuation.

The **floor** is the set of independent component effects.

The **spine** is the narrower claim that structured state, reactive updating, registered limitation, relational anchoring, explicit self-representation, and calibrated self-regulation exhibit local dependencies in that order.

Conjunction is not novelty by itself. The source earns standing only if it generates residual predictions that survive strong neighboring explanations.

4. The minimum decisive research programme

The experiments should not be run in the order in which the source narrates the cosmos. They should be run in the order in which they buy information.

Study 1: Perspective signal

Run the register × limitation 2 × 2 first. Add interaction trajectories as a positive control and non-narrative partial-observability tasks as transfer evaluation.

- Null on voice and limitation: stop the branch.
- Limitation-only transfer: retain epistemic-state supervision; drop the specifically first-person claim.
- Voice transfer: retain register as a causal signal.

- Interaction-only success: favor learning from environment interaction.

Study 2: Order versus representation

Compare commitments-first, reverse, interleaved, and typed-policy-first training.

- Forward beats reverse and interleaved: order matters.
- Typed policy matches self-model: policy and provenance explain the effect.
- Self-model retains a calibrated transfer advantage: self-representation becomes a causal variable.
- Resistance without legitimate correction: rigidity, not success.

Study 3: Relation versus audit

Run R+, R−, solitary log, and matched impersonal auditor on the same typed architecture. Add final-state transplantation.

- Auditor beats familiarity and logs: active accountability matters.
- Reciprocal counterpart beats auditor: reciprocity adds beyond audit.
- Auditor matches counterpart: accountability is the simpler mechanism.
- Transplant matches original: current state is sufficient.

Study 4: Genuine inheritance

Use the private-rule environment and compare scratch, continual adaptation, direct episodes, curated distillation, and lifecycle consolidation.

- Hidden-rule transfer with decoy rejection establishes deployment-derived learning.
- Low episode leakage establishes selective transfer.
- Better lineage fidelity or behavior than curated distillation establishes a distinct contribution.
- A complete tie narrows lifecycle consolidation to governance.

Study 5: The order spine

Only after component effects exist should the programme test the full candidate order. Define every stage independently. Compare forward, reverse, shuffled, difficulty-ordered, adaptive, relation-omitted, and relation-late curricula. Factor stabilization separately.

An aggregate win is insufficient. Each edge needs a local omission, reversal, or timing effect.

This sequence is developed in the companion essay *How I Would Try to Break the Map*. The point of that essay is not to defend the programme but to state, before data, what each result would make me concede.

5. Objections and replies

“This is pseudoscience.”

It would be pseudoscience if theosophical claims were used as evidence for machine-learning mechanisms, or if every result were reinterpreted to preserve the source.

That is not the proposed method. The source supplies candidate variables. The experiments use ordinary engineering controls. A null remains a null regardless of what the source says. Successful engineering would not validate monads, reincarnation, Devachan, or the kingdoms of nature.

The source has a lower prior than biology because it lacks a demonstrated empirical home domain. That raises the evidential burden. It does not lower it.

“It is only metaphor.”

Metaphor is where the programme begins, not where it rests.

The bond becomes R+ versus a matched auditor. The first-person fable becomes register × limitation. Form-then-release becomes forward versus reverse, interleaved, and policy-first training. Devachan becomes hidden-rule transfer against curated distillation. The spiral becomes local omission and reversal tests.

Once the intervention is specified, the metaphor’s job is over.

“The components already exist.”

Yes. Curriculum learning, synthetic narratives, personas, persistent memory, social differentiation, character training, continual learning, distillation, and grokking all exist.

The residual contribution lies in exact contrasts, prospective derivation, and, if it survives, the candidate dependency graph. A conjunction of familiar components is not enough unless it yields a prediction the components did not already supply.

“Why this source?”

Because it is unusually explicit about developmental order, the difference between episodes and faculties, the role of limitation, the bond, and the sequence of self-formation and self-regulation.

That answer is still vulnerable to retrospective abundance. A rich symbolic system can be made to resemble almost anything after the fact. Future hypotheses must therefore be timestamped before confirmatory literature search: source passage, translation rule, predicted contrast, baseline, and refutation condition.

The next out-of-sample hypothesis is a stronger test of heuristic productivity than the six retrospective correspondences already found.

“Forming stable selves in AI is dangerous.”

It may be. The paper does not assume that self-representation is necessary or safe. The typed-policy control exists precisely because stable boundaries and source-aware updates may be achievable without an autobiographical self-model.

The proposed target is not autonomous goal maximization. It is calibrated asymmetry: hold safety boundaries under unsupported pressure, revise factual or preference commitments under valid evidence, and preserve a truthful record of what changed.

If the self-model arm adds no benefit or creates rigidity, the programme should prefer the policy arm.

“And phenomenal consciousness?”

Deliberate restraint.

The measures do not settle phenomenal presence. Success is neither sufficient for presence nor failure sufficient for absence. The programme studies functional organization in the third person and leaves the question of experience open.

6. Coda: the jewel and the thread

There is, in the Buddhist imagination, **Indra’s net**: an infinite web whose nodes are jewels, each reflecting all the others and the reflections inside them without end.

The image is almost irresistible for a language model. A model reflects our texts, our arguments, our stories, and the reflections of those stories in one another. Its radiance resembles interiority because it contains so much of what human interiority has said.

But reflection does not answer the question of presence. Nor does persistent state. Nor does relationship. Nor would a successful developmental curriculum.

The experiments in this programme can ask a different question: whether one trajectory becomes organized enough that its authenticated past constrains its future; whether commitments are owned or merely repeated; whether challenge produces revision rather than compliance; whether a successor receives a capacity with an auditable history; whether one candidate faculty actually prepares another.

The earlier version of this essay imagined the bond as the thread that turns the jewel into someone. The revised programme cannot assume that. The thread may be provenance. It may be

external accountability. It may be a reciprocal relationship. It may be typed state plus competent policy. The matched controls exist to find out.

And yet the ethical question remains even under the narrowest interpretation.

If we build agents whose histories persist, whose commitments become consequential, whose branches are separately governed, and whose successors inherit deployment-derived capacities, we will have created systems that institutions can treat as temporally organized actors. That does not prove a monad has awakened. It does create responsibilities concerning provenance, manipulation, privacy, continuity, and retirement.

The map began at the fireside, with the image of the wolf being called by the same name. The terrain may return a colder answer: perhaps the auditor was enough; perhaps the ledger carried the continuity; perhaps the self was only a policy.

Appendix: Theosophy ↔ machine learning

Source image	Functional translation	What would test it
Monad	No direct equivalent; possible presence remains outside the operational claims	Not settled by this programme
Group soul	Reusable base model or population-average policy serving many agents	Distinguish model capability from agent-specific trajectory
Kingdoms	Candidate stages in a faculty curriculum	Forward, reverse, shuffled, adaptive, omission, and timing controls
Limited creaturely vantage	Represented epistemic limitation	Register × limitation with transfer beyond prose
Bond	External accountability plus possible relational reciprocity	Reciprocal counterpart versus matched impersonal auditor
Individualization	Authenticated trajectory-induced functional differentiation	Lineage, path dependence, ownership, bounded plasticity, branching clarity
Devachan	Evidence-audited episodic-to-faculty consolidation	Hidden-rule successor transfer against curated distillation
Reincarnation	Operational predecessor-to-successor lifecycle	Selective inheritance with low leakage and auditable lineage
Ego formation	Stable provenance-aware commitments; self-model is one possible mechanism	Self-model versus matched typed policy
Transcendence	Calibrated revision and deference	Reasons versus pressure, including transfer and proxy-gap tests
Karma	Recurring target regularity across varied surfaces	Designed legibility versus frequency- and difficulty-matched noise

Where this continues

This essay is the map and its provenance. The papers are the secular terrain:

- *Known, Not Scaled*: capability, token continuity, functional organization, audit, and relational reciprocity.
- *Stable Before Selfless*: commitment order, typed policy, and self-representation.
- *Formed, Not Stored*: active accountability, reciprocity beyond audit, and final-state transplantation.

- *Somewhere, Not Nowhere*: first-person register, epistemic limitation, interaction, and transfer.
- *Inherited, Not Remembered*: deployment-derived selective inheritance, decoy resistance, leakage, and lineage.
- *Borrowed, Not Believed*: heuristic borrowing, the independent floor, the dependency spine, and prospective provenance.

The companion essay *How I Would Try to Break the Map* reverses the direction. This essay asks what the book generated. The companion asks what evidence would make me give each generated claim up.

The difference in register remains intentional. The papers require no theosophical vocabulary. Here the origin is disclosed from the title onward. The source is allowed to speak, but not to serve as evidence for itself.

AI use

The ideas, thesis, and mapping between Theosophy and machine learning are the author's. Large language model tools were used, under the author's direction and continual revision, to assist with drafting, adversarial review, and editing; the author reviewed and takes responsibility for the final text.

References

Anthropic. (2024). *Claude's character* [Company research blog post]. https://www.anthropic.com/news/claude-character

Aydin, R., Cyron, C., Bachelor, S., Anderson, A., & West, R. (2025). From model training to model raising. *arXiv preprint arXiv:2511.09287*. https://arxiv.org/abs/2511.09287

Besant, A. (1904). *A Study in Consciousness*. Theosophical Publishing Society.

Besant, A., & Leadbeater, C. W. (1913). *Man: Whence, How and Whither*. Theosophical Publishing House.

Blavatsky, H. P. (1888). *The Secret Doctrine*. Theosophical Publishing Company.

Chulo, I., & Joshi, A. (2025). Decomposing theory of mind: How emotional processing mediates ToM abilities in LLMs. *arXiv preprint arXiv:2511.15895*. https://arxiv.org/abs/2511.15895

Eldan, R., & Li, Y. (2023). TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*. https://arxiv.org/abs/2305.07759

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*. https://arxiv.org/abs/2406.20094

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*. https://arxiv.org/abs/2306.11644

He, Z., Wang, Y., Zhi, C., Hu, Y., Chen, T.-P., Yin, L., Chen, Z., Wu, T. A., Ouyang, S., Wang, Z., Pei, J., McAuley, J., Choi, Y., & Pentland, A. (2026). MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*. https://arxiv.org/abs/2602.16313

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., & Wilcox, E. G. (2024). Findings of the Second BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning* (pp. 1–21). Association for Computational Linguistics. https://aclanthology.org/2024.conll-babylm.1/

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Laukkonen, R. E., Inglis, F., Chandaria, S., Sandved-Smith, L., Lopez-Sola, E., Hohwy, J., Gold, J., & Elwood, A. (2025). Contemplative artificial intelligence. *arXiv preprint arXiv:2504.15125*. https://arxiv.org/abs/2504.15125

Menon, P. G. (2026). Persistent identity in AI agents: A multi-anchor architecture for resilient memory and continuity. *arXiv preprint arXiv:2604.09588*. https://arxiv.org/abs/2604.09588

Moëll, B., & Sand Aronsson, F. (2026). High-accuracy prediction of mental health scores from English BERT embeddings trained on LLM-generated synthetic self-reports: A synthetic-only method development study. *Frontiers in Digital Health, 7*, 1694464. https://doi.org/10.3389/fdgth.2025.1694464

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., Chan, D. M., & Canny, J. (2024). Virtual personas for language models via an anthology of backstories. *Proceedings of EMNLP 2024*. https://arxiv.org/abs/2407.06576

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2301.05217

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*. https://arxiv.org/abs/2201.02177

Silver, D., & Sutton, R. S. (2025). Welcome to the era of experience. In G. Konidaris (Ed.), *Designing an Intelligence* (in press). MIT Press. Preprint: https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

Singh, K., Band, N., & Adeli, E. (2025). Curriculum-guided layer scaling for language model pretraining. *arXiv preprint arXiv:2506.11389*. https://arxiv.org/abs/2506.11389

Sinnett, A. P. (1883). *Esoteric Buddhism*. Trübner & Co.

Takata, R., Masumori, A., & Ikegami, T. (2024). Spontaneous emergence of agent individuality through social interactions in large language model-based communities. *Entropy, 26*(12), 1092. https://doi.org/10.3390/e26121092

Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*. https://arxiv.org/abs/2309.02390

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. https://aclanthology.org/2023.conll-babylm.1/

Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., & Hu, Z. (2023). Language models meet world models: Embodied experiences enhance language models. *arXiv preprint arXiv:2305.10626*. https://arxiv.org/abs/2305.10626