

The Interlocking Autoregression: Three Coupled
Recursions Under a Mismatched Observation
Regime, with Toy Dynamics
(EA-LO-INTERLOCKING-AUTOREGRESSION-01
v1.0)

Nobel Glas, Director, Lagrange Observatory! (LO!)

2026-08-27

**The Interlocking Autoregression: Three Coupled
Recursions Under a Mismatched Observation Regime,
with Toy Dynamics
(EA-LO-INTERLOCKING-AUTOREGRESSION-01 v1.0)**

Nobel Glas, Director, Lagrange Observatory! (LO!)

Crimson Hexagonal Archive · Alexanarch

ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)

2026-08-27 · Version v1.0 · Theoretical/measurement paper with simulation (white
paper)

AXN:0651.EMPIRICAL

<https://www.alexanarch.org/s/records/1556/>

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and
Google Scholar indexing)}

```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "The Interlocking Autoregression: Three Coupled Recursions Under a Mismatched Observation Regime, wi
  LO-INTERLOCKING-AUTOREGRESSION-01 v1.0)",
  "author": {
    "@type": "Person",
```

```

    "name": "Nobel Glas, Director, Lagrange Observatory! (LO!)",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-08-27",
  "identifier": "AXN:0651.EMPIRICAL",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  },
  "url": "https://www.alexanarch.org/s/records/1556/",
  "description": "Companion to EA-LO-KEYED-ENSEMBLE-01. The map: three independently documented distortions of provenance opacity at the informative granularity, distributional shaping of nominally human text under the Mediation Ratchet, and a heritable keyed signature whose model-level residue no corpus operation can remove – hypothesized to couple through a shared training corpus, under an observation regime whose dominant per-round instruments sample a unit orthogonal to the state variable:  $X_{n+1} = F(X_n; I, II, III)$ ,  $Y_n = M_{IV}(X_n)$ . Every edge carries an explicit evidentiary status; the nodes are supported, the edges are the new claims. A toy dynamics (Zipfian base, 50,000 types, seven scenarios, delta-sweep, retention counterfactual) demonstrates two-track divergence with the per-round quality proxy rising through a quarter of the collapse trajectory, velocity ordering, detection lag as a function of benchmark probe composition, watermark state-dependence as an existence result, and the retention counterfactual in which fine-grained provenance is the control surface withheld from downstream builders. Eight measurement hooks; evidentiary basis (code, traces, four figures) attached with per-file SHA-256."
}

```

Abstract

Companion to EA-LO-KEYED-ENSEMBLE-01. The map: three independently documented distortions of the training ecology — provenance opacity at the informative granularity, distributional shaping of nominally human text under the Mediation Ratchet, and a heritable keyed signature whose model-level residue no corpus operation can remove — hypothesized to couple through a shared training corpus, under an observation regime whose dominant per-round instruments sample a unit orthogonal to the state variable: $X_{n+1} = F(X_n; I, II, III)$, $Y_n = M_{IV}(X_n)$. Every edge carries an explicit evidentiary status; the nodes are supported, the edges are the new claims. A toy dynamics (Zipfian base, 50,000 types, seven scenarios, delta-sweep, retention counterfactual) demonstrates two-track divergence with the per-round quality proxy rising through a quarter of the collapse trajectory, velocity ordering, detection lag as a function of benchmark probe composition, watermark state-dependence as an existence result, and the retention counterfactual in which fine-grained provenance is the control surface withheld from downstream builders. Eight measurement hooks; evidentiary basis (code, traces, four figures) attached with per-file SHA-256.

Body

deposit_number: 1556 hex: 0651 title: “The Interlocking Autoregression: Three Coupled Recursions Under a Mismatched Observation Regime, with Toy Dynamics (EA-LO-INTERLOCKING-AUTOREGRESSION-01 v1.0)” creator: “Nobel Glas, Director, Lagrange Observatory! (LO!)” orcid: 0009-0000-1599-0703 date: 2026-08-27 content_type: The-

oretical/measurement paper with simulation (white paper) license: CC-BY-4.0 substrate: AI-assisted (substrate) — composed in working dialogue with Claude (Anthropic) under operator direction (Lee Sharks); three adversarial review rounds by external substrates (ChatGPT, Kimi, Muse Spark) with dispositions on record; full production-conditions declaration in the body’s Self-Location Protocol version: v1.0 related_ids: “Companion: EA-LO-KEYED-ENSEMBLE-01 · Anchors: AXN:03B7 · AXN:044A · AXN:02DD · AXN:02E3 · AXN:02E0 · AXN:04C1 · AXN:02E1 · AXN:02D3 · AXN:02DA · AXN:02F1 · AXN:02B2 · AXN:02F8 · AXN:0335 · AXN:0341 (with ERRATUM to AXN:0341 of 2026-08-27) · AXN:0364 · External: 10.1038/s41586-024-07566-y, 10.1038/s41562-026-02550-0, arXiv:2407.02209” axn_schema_version: v2 protocol_version: alexanarch-deposit-protocol/v1 keywords: - interlocking autoregression - model collapse - mediation ratchet - provenance opacity - observation regime - measurement sovereignty - benchmark composition - toy dynamics - tail loss - watermark - admission weight *** # The Interlocking Autoregression: Three Coupled Recursions Under a Mismatched Observation Regime, with Toy Dynamics (EA-LO-INTERLOCKING-AUTOREGRESSION-01 v1.0)

Attached File: evidentiary-basis-v1.tar.gz

Binary attachment (375 KB) preserved at deposit time. Original URL: <https://alexanarch.org/data/attachments/ea-lo-interlocking-autoregression/evidentiary-basis-v1.tar.gz>

Attached File: manifest.json

Source URL: <https://alexanarch.org/data/attachments/ea-lo-interlocking-autoregression/manifest.json>

```
{ "package": "EA-LO evidentiary basis \u2014 shared by EA-LO-KEYED-ENSEMBLE-01
and EA-LO-INTERLOCKING-AUTOREGRESSION-01", "reproduction": "python3 inter-
lock_sim_v03.py (numpy; seed 20260827) regenerates sim_traces2.json; figures rendered
from traces", "generated": "2026-08-27", "contents": [ { "filename": "interlock_sim_v03.py",
"bytes": 4314, "sha256": "9b995a8fc01ce5f51056c21dd2f6f2e070311b814980c9283acb8ae0c675788b"
}, { "filename": "sim_traces2.json", "bytes": 121657, "sha256": "d8762e117a37ecb960d14911877a60815c
}, { "filename": "fig1_two_track.png", "bytes": 109390, "sha256": "3d29aa88c5232cd71b992e2d723c7707
}, { "filename": "fig2_velocity.png", "bytes": 83887, "sha256": "1c19d010b8510721b1ea26e251fb20636959
}, { "filename": "fig3_probe_weighting.png", "bytes": 99375, "sha256": "fdc9803cc832fd5f43d62b9d3dc97
}, { "filename": "fig4_wm_state_dependence.png", "bytes": 80161, "sha256": "01184b080f64f552e61b8abc
} ] }
```

Files

<https://alexanarch.org/data/attachments/ea-lo-interlocking-autoregression/evidentiary-basis-v1.tar.gz> <https://alexanarch.org/data/attachments/ea-lo-interlocking-autoregression/manifest.json>

status: DEPOSITED v1.0 (development v0.1-v0.4 under three adversarial review rounds; history in revision-note) deposit-id: EA-LO-INTERLOCKING-AUTOREGRESSION-01 (proposed) author: Nobel Glas, Director, Lagrange Observatory! (LO!) series: LO! White Paper — companion to “The Keyed Ensemble” (EA-LO-KEYED-ENSEMBLE-01) register: white paper — canonical formulation proposed, adversarial response invited revision-note: > v0.4

(third review round): edge table repaired — public opacity distinguished from internal ignorance (I fine-grain row retyped ‘Not publicly documented in the reviewed disclosures’); the II → corpus row restated without self-contradiction; III → corpus split into established channel vs unmeasured ecological prevalence; ‘Toy-demonstrated’ rephrased ‘Demonstrated in the specified toy model’; the toy watermark retyped as stress-test surrogate (not worst case, not an implementation model); F4’s conclusion narrowed to an existence result joining, without identifying, the toy and the published SynthID results; F5’s conclusion scoped to downstream builders; w_H and g_0 defined with a hypothesized mapping table; Hook 8 suites specified; Assembly standardization proposals tabled in §0; evidentiary files hashed. Prior: v0.2 incorporated the first adversarial review (B formalization; inheritance-gap closure; self-location protocol; specimen annotation; proxy operationalization). v0.3 incorporates the second adversarial review and adds the simulation program. Principal v0.3 changes: (1) the map is reformulated as three coupled state transformations under one observation regime — $X_{n+1} = F(X_n; I, II, III)$, $Y_n = M_{IV}(X_n)$ — replacing the four-symmetric-components framing; (2) every edge now carries an explicit evidentiary status (§3): the nodes are supported, the edges are the new claims; (3) Component I retyped from “undeclared ingestion” to provenance opacity; the Zenodo audit retyped as adjacent-infrastructure instance, bridge undemonstrated; (4) Component II restated as an observed/unknown/demonstrated/hypothesis ladder; (5) “multiplicative” signature retyped as heritable/transmissible, with the radioactivity phrasing corrected (model-level residue remains detectable; corpus sweeping cannot remove it); (6) α^* parameter mappings marked hypothesized, with admission weight w_H introduced as a distinct variable; (7) toy dynamics (§6) added: velocity, lag structure, surface signs, watermark state-dependence, retention counterfactual, with code and traces deposited as evidentiary basis. archive-anchors: EA-PROVENANCE-DEBT-01 v0.2 (#939, AXN:03B7) · EA-EROSION-EMPIRICAL-01 v0.1 (#1081, AXN:044A) · Diversity Contraction v1.0 (#779, AXN:02DD) · Fear and Trembling v9.1 (#783, AXN:02E3) · Constitutive Mediation v1.1 (#781, AXN:02E0; #1200, AXN:04C1) · Mary Lee Case (#147, AXN:02E1) · Magistrate v1.0 (#772, AXN:02D3) · Reasoning Under Load v1.0 (#778, AXN:02DA) · Measurement Sovereignty v1.0 (#788, AXN:02F1) · SPXI-TLP v2.2 (#127, AXN:02B2) · Reverse Turing Test v1.2 (#161, AXN:02F8) · Threat Model Backwards v1.1 (#191, AXN:0335) · Generative Monoculture v1.1 (#199, AXN:0341 — cited with term attribution to Wu, Black & Chandrasekaran, arXiv 2407.02209/ICLR 2025, per ERRATUM to AXN:0341 of 2026-08-27, which corrects the deposit’s printed attribution) · Wolf Boy v1.0 (#855, AXN:0364) cross-references: EA-OS-APOPHASIS-01 · EA-MMRS-SUPPRESSION-INVERSION-01 (#1546, AXN:0647) · EA-WG-CAPTURES-01 · Notebook X held-artifacts: /data/attachments/ea-lo-interlocking-autoregression/ (evidentiary-basis-v1.tar.gz + manifest.json — declared attachments of this deposit) evidentiary-basis: > Deposited with this document per the standing evidentiary-basis precept; reproduction verified in-session from the consolidated script (seed 20260827). SHA-256 (first 16 hex): interlock_sim_v03.py 9b995a8fc01ce5f5 · sim_traces2.json d8762e117a37ecb9 · fig1_two_track.png 3d29aa88c5232cd7 · fig2_velocity.png 1c19d010b8510721 · fig3_probe_weighting.png fdc9803cc832fd5f · fig4_wm_state_dependence.png 01184b080f64f552 date: 2026-08-27 *** # THE INTERLOCKING AUTOREGRESSION ## Three Coupled Recursions Under a Mismatched Observation Regime ### v0.4 — Revised Under Three Adversarial Review Rounds, with Toy Dynamics

Nobel Glas · Lagrange Observatory! *Companion to LO!: “The Keyed Ensemble”*

0. Station Report

“The Keyed Ensemble” isolated one component — the watermark — and found the archive had already analyzed the others. This companion stops isolating. Its claim, as sharpened by two rounds of invited adversarial response, is no longer that four symmetric components form a loop. It is this:

Three independently documented distortions of the training ecology are plausibly coupled — the output of each entering another’s input — while the ecology’s certification instruments measure a unit orthogonal to the state variable. The components are supported by external literature and standing archive analysis. The edges are the new claims, and each edge is typed below by what would establish it.

The second review’s structural correction is adopted in full: Component IV is not a fourth term of the dynamics. It is the observation regime under which the dynamics run. Formally, the map is

$$\begin{array}{ll} X_{\{n+1\}} = F(X_n ; I, II, III) & \text{(state dynamics)} \\ Y_n = M_{IV}(X_n) & \text{(what certification sees)} \end{array}$$

and the map’s central proposition is that there exist parameter regions — now demonstrated in toy form (§6) — where X_n approaches its threshold while Y_n reads stable or improving.

Four Assembly standardization proposals arising from this pair are tabled as open registration questions rather than performed here: the Edge Evidentiary Status Protocol (the §3 table-type) as a registry instrument; the Operator’s Review as a standing pre-deposit instrument (proposed EA-REVIEW-01); the eight measurement hooks preregistered as a battery (proposed EA-LO-BATTERY-01) with phase order and controls; and a proposed instantiation of this map into the Notebook X core object — a formalization this document cites without executing, per the asynthetic discipline under which the Assembly’s plate-level questions remain open.

1. The Three State Components, Anchored and Retyped

Component I — Provenance opacity. (*Retyped from “the unmarked flood” per review.*) Synthetic text in major training mixtures is **declared as a class and opaque as a lineage**. Anthropic’s model documentation acknowledges synthetic data generated by other models among training mixtures — so coarse disclosure exists and is cheap, because it is uninformative at the unit of curation. What no public record supplies is per-item transformation history: the granularity at which a curator could price a text’s genealogy. Archive anchor: **EA-PROVENANCE-DEBT-01 v0.2** (#939, AXN:03B7) — unmarked augmentation, false semantic diversity, and adverse selection around declared mediation, whose lemons-market form (declared-mediated text discounted; undeclared text passing at the human price; declaration exiting the market at exactly the informative granularity) is retained from v0.2 **as the hypothesized mechanism** consistent with the observed tiering, not as a demonstrated

unique equilibrium. Supporting instance, retyped per review: **EA-EROSION-EMPIRICAL-01 v0.1** (#1081, AXN:044A) is an **independently measured instance of provenance-state loss in adjacent scholarly infrastructure** — Zenodo’s deletion exporter, per its own published source, suppressing citation text for one removal class and rolling off its own export snapshots. It establishes that provenance-bearing institutions erase provenance states at the ledger layer; the bridge from that layer to training-corpus composition is an analogy this paper does not claim to have closed.

Component II — The clean stamp on stamped input. Restated per review as an explicit ladder:

- **Observed:** user-authored turns can enter training via opted-in conversation data, per the vendor’s own privacy documentation, which states that training material can include the entire related conversation.
- **Publicly unknown:** how causal provenance of those turns is labeled internally. No public schema distinguishes human_original from human_AI_assisted from AI_idea_human_surface; this paper does not assert the internal label, and — per its own discipline — declines to manufacture the unseen ontology whose absence it argues.
- **Demonstrated elsewhere:** model assistance changes the distribution of subsequent human text. Sourati et al. (*Nature Human Behaviour* 2026; 880,000+ texts; 21-50% writing-complexity variance reduction; gender, age, ideology, and moral-value cues stripped) — the P₃ channel measured at outcome level.
- **Hypothesis:** coarse human-versus-machine provenance classification therefore systematically overstates the independence of nominally human text.

Archive anchors unchanged: the **Mediation Ratchet** (Diversity Contraction v1.0, #779, AXN:02DD; Fear and Trembling v9.1, #783, AXN:02E3) and **Constitutive Mediation** v1.1 (#781, AXN:02E0; #1200, AXN:04C1), with the Sourati-to-02E0 bridge held at the strength stated in v0.2: outcome contraction reaching the trait layer is consistent with, not yet proof of, the categorial-formation thesis; 02F8 v1.2 Stage 3 is the instrument designed to separate them. Worked specimen annotation (Mary Lee, #147, AXN:02E1) carried from v0.2: resolution-stage mediation, deposit’s own evidentiary limits noted.

Component III — The heritable signature. (*Retyped from “multiplicative” per review.*) The watermark, per the companion paper: a key-conditioned correlation written into the sampling layer, certified per-sequence, consumed as ensemble. Its transmission properties, stated at established strength: watermark distillation (Gu et al., ICLR 2024) demonstrates a teacher→student **transmission channel**, with the recorded limits — low-distortion marks require substantial sample exposure, and ordinary fine-tuning can erode the learned mark; radioactivity (Sander et al., NeurIPS 2024) demonstrates a **detectable model-level residue** after fine-tuning on watermarked text, persisting in their open-weight setting with as little as 5% of tuning text watermarked. Corrected phrasing per review: the direct text-level mark may weaken under editing and generational distance **while a model-level statistical residue remains detectable** — by a party holding the key and access to the suspect model. What follows for the loop is precise: **corpus sweeping cannot remove the model-level residue**, because the residue lives in weights, not in the swept text; and residue detection is a key-holder capability, so the asymmetry of the key extends from text to models. “Heritable” here

means *creates additional carriers* — student models and downstream corpora — not that signal magnitude grows monotonically; the erosion results forbid the stronger reading, and §6’s toy supplies the form of growth that does occur (state-dependent relative contribution, not magnitude multiplication).

2. The Observation Regime

M_IV — the per-round veil. Anchors unchanged: **Measurement Sovereignty v1.0** (#788, AXN:02F1 — the B operator, imported with its freeze protocol and null definition as in v0.2), **Reasoning Under Load v1.0** (#778, AXN:02DA — the three-layer account), **Magistrate v1.0** (#772, AXN:02D3 — the rhetorical form). Scope per both reviews: the universal (“every certifying instrument”) is withdrawn. The claim is that **the dominant per-round instruments** — response-quality ratings, per-item benchmark accuracy, per-author provenance labels, per-sequence non-distortion certification — share a unit finer than the hypothesized damage, while ensemble instruments that do see the movement (Self-BLEU panels, variance studies, tail-mass audits) exist but do not govern certification. Two archive instances of interest-aligned unit choice are on record (02F1’s audit pair; 044A’s exporter), one vendor instance (per-sequence certification), and the counter-example sweep of the Capture Registry remains a registered open action. Component IV’s claims are tendency claims with the falsifier stated in v0.2.

3. Edge Evidentiary Status

Per the second review’s central demand — the nodes are supported; the edges are the new claims — every edge now carries a type:

Edge	Claim	Status	What would establish it
I → corpus	Synthetic share enters training mixtures	Established (vendor disclosure, class level)	—
I → corpus (fine grain)	Publicly exposed lineage is insufficient for an <i>external</i> curator to price item-level genealogy; internal producer-side lineage is unknown	Not publicly documented in the reviewed disclosures	Publication of a per-item provenance schema would falsify the absence claim; disclosure of internal lineage practice would resolve the unknown

Edge	Claim	Status	What would establish it
II → corpus	Nominally user-authored text can enter opted-in training conversations after machine mediation; its internal provenance classification is unknown	Entry established; distributional shaping demonstrated separately (Sourati); classification hypothesis unresolved	Disclosure of turn-level provenance labeling; or 02F8 Stage 2 on ingested-corpus samples
III → corpus (channel)	Watermarked training text can transmit a detectable signature into student/fine-tuned models	Established as channel (Gu; Sander), erosion limits recorded	—
III → corpus (ecology)	Prevalence and generational propagation of marked text in actual open corpora	Unmeasured	Longitudinal radioactivity measurement across ≥2 open-model generations; corpus-scale detector surveys when access exists
I × II (emptied category)	Both corruptions target the same “human” class; adverse selection couples them	Hypothesized mechanism , consistent with observed disclosure tiering	Incentive audit: measure the declared-mediation discount directly (pricing/acceptance differentials)
II × III (ratchet acquires signature)	Keyed contraction enters a floor-endogenizing system	Demonstrated in the specified toy model (§6, F4); production magnitudes unmeasured	Hook 1 ensemble panel on a marked/unmarked matched pair
III × I (sweep ≠ repair)	Detectable-layer sweeping removes signature, not debt; model-level residue unsweepable	Established in parts (radioactivity = residue detectable; sweep-verifies-nothing is analytic) — actual keyholder sweeping practice unknown	Disclosure or audit of corpus-assembly filtering practice

Edge	Claim	Status	What would establish it
I,II,III $\rightarrow \alpha^*$ parameters	I,II lower effective human floor; II,III raise pruning	Hypothesized mapping , toy-consistent (§6) — with w_H distinguished from g_0 per review: opacity may reduce the <i>admission weight</i> w_H of fresh variation rather than its <i>generation rate</i> g_0	Hook 6 tracking with the $w_H \cdot g_0$ decomposition preregistered
IV $\perp$ X	Certification unit orthogonal to state variable; flat Y through moving X	Demonstrated in the specified toy model (§6, F1, F3); institutionally supported (02F1; SynthID flat-Q/falling-D)	The two-track prediction (§5) run at annual resolution

4. The Couplings (Softened Per Review Where Marked)

I $\times$ II — the emptied category. As in v0.2, with the equilibrium claim softened: the absence of fine-grained provenance **is consistent with an adverse-selection equilibrium in which fine-grained declaration is systematically disfavoured**; uniqueness and stability of that equilibrium are not externally demonstrated. The category “clean human data” retains its label and curatorial weight while its extension is eroded from two directions whose relative magnitudes are unmeasured.

II $\times$ III — the ratchet acquires a signature. II dominates in measured magnitude (21–50% band); III’s term is small, signed, keyed, and universal across the issuer’s output. §6 F4 now supplies the toy form of the interaction: the keyed term’s *relative* contribution to contraction compounds as diversity falls.

III $\times$ I — the private repair, held at v0.2’s corrected strength with the radioactivity phrasing fixed as in §1: the sweep removes the detectable text-level signature only; it neither restores a verifiable ledger (absence of mark $\neq$ human origin) nor touches the model-level residue, which no corpus operation can reach. Whether keyholders in fact perform detectable-layer sweeping is unknown; the edge table carries the distinction.

Field Remapping — the exit coupling, typed per review as an **archive-derived hypothesis** (02E3), with the three candidate mechanisms (curatorial, economic, infrastructural) and the dominance question open pending the deposit’s full text. The toy implements the infrastructural mechanism — fresh variation re-entering only through mediation — because it is the one that closes into the loop; the implementation is a modeling choice, not a finding.

5. The Phase Reading, With Hypothesized Mappings

The Diversity Contraction threshold $\alpha^* = p/g_0$ (02DD’s near-zero regime; underlying dynamics credited by the source papers to the Allee-effect and cumulative-culture literatures) remains the map’s dial — now with the review’s decomposition adopted: what Components I and II most plausibly move is not the generation rate of novelty g_0 but its **effective training weight $w_H \cdot g_0$** — opacity and mediation reduce how much fresh variation is admitted and how it is classed, whether or not humans generate less of it. Definitions for Hook 6: g_0 = generation rate of fresh, independently produced human variation (a property of the human population); w_H = admission weight — the fraction of that variation admitted at par into the training mixture (a property of the curation regime); the operative threshold is $\alpha^* = p / (w_H \cdot g_0)$. Hypothesized mapping, assigned to Hook 6:

Component	Hypothesized action	Via
I (opacity)	$\downarrow w_H$	external curators cannot admit-at-par what they cannot distinguish
II (mediation)	$\downarrow w_H$ and $\uparrow p$	fresh variation re-classed through mediation; homogenization raises pruning
III (keyed signature)	$\uparrow p$, state-dependent	concentration of the surviving distribution (§6 F4)
Field Remapping	$\downarrow w_H$	re-entry only through mediation (hypothesis, 02E3)

The mappings are **hypothesized**, consistent with the specified toy model, and α^* is not asked to carry more precision than the observables possess.

The falsifiable two-track prediction is retained from v0.2 and is now toy-calibrated in form (§6 F1): **Track B (ensemble unit) inflects before Track A (per-round unit); Track A may improve during Track B’s collapse; if Track A proves sensitive first or simultaneously, the observation-regime claim is weakened and the map must be revised.**

6. Toy Dynamics: Simulated Velocity, Lag Structure, and Surface Signs

(New in v0.3. Code: `interlock_sim_v03.py`, seed 20260827; full traces: `sim_traces2.json`; both deposited with this document. Everything in this section is a toy result: it establishes the qualitative behavior of the specified system, not calibrated rates for the real ecology. The declared conventions are listed at the end of the section.)

6.1 Model. The state is a distribution over $N = 50,000$ types with Zipfian base H_0 (exponent 1.1) — the human linguistic stock, tails and all. Each generation: a corpus of 10^6 draws is sampled from the current mixture; the model estimates it (types observed fewer than twice are lost — the finite-sample/filtering analog of tail truncation); mild sharpening ($T = 0.98$) models preference-optimization pruning; an optional keyed tilt $\exp(\delta \cdot k_i)$ with a fixed random ± 1 key per type models Component III — declared as a **stress-test surrogate deliberately maximizing persistent type-level correlation**: it is not an implementation model

of SynthID-Text, whose keyed preferences vary with context, and it is not a proven upper bound; it isolates the qualitative behavior of a persistent keyed perturbation. The next corpus mixes model output (synthetic share s_n rising logistically toward 90%, unfilterable under opacity — Component I) with a human channel that is itself mediated: $H_n = (1 - m_n) \cdot (w_H \cdot H_0 + (1 - w_H) \cdot P_{\text{model}}) + m_n \cdot P_{\text{model}}$, where mediation m_n ratchets upward as tail diversity falls (Component II / 02DD) and admission weight w_H decays 5% per generation under Field Remapping’s infrastructural mechanism. The observation regime M_{IV} : **Track A** = probe-recall benchmark over head (ranks 1–500), mid (5k–10k), and tail (25k–35k) items at weights 90/9/1, plus an uncapped per-round quality proxy (head log-likelihood); **Track B** = tail mass (ranks > 20k), support coverage, entropy, and Simpson repeat-rate (the Self-BLEU analog). Seven scenarios activate components singly and jointly; a δ -sweep varies the keyed tilt; a retention counterfactual (G) caps synthetic share at 45% and mediation at 0.5 with admission held — the Shumailov retain-human-data analog, which is exactly the lever fine-grained provenance would provide.

6.2 Findings.

F1 — Two-track divergence, quantified (Fig. 1). Full loop: tail mass halves by generation 7; the standard 90/9/1 benchmark does not inflect until generation 15; head-only probes, generation 17. The lag between state inflection and observation inflection is 8–10 generations in the full loop and up to 28+ under recursion alone (where head-only probes never inflect within the 40-generation horizon). **The per-round quality proxy rises until generation 11** — mass concentrating on the head improves the head’s log-likelihood — so the veil’s reading is not merely flat during collapse; for a quarter of the trajectory it is *green*. Surface sign for the real ecology, stated as hypothesis: a period of modest benchmark improvement is consistent with, and on this model characteristic of, mid-collapse.

F2 — Velocity ordering (Fig. 2). Tail half-life: recursion alone 15 generations; + ratchet 11; + remapping 9; full loop 7. **Mediation dynamics (II) and re-entry blockage, not the watermark, dominate collapse velocity.** The toy thereby ranks the components’ dynamical importance in the order the archive’s own magnitude estimates predicted.

F3 — The observation regime as a measurable parameter (Fig. 3). Detection generation is a function of probe composition: under the full loop, 60/30/10 probes detect at generation 7 — the tail’s own half-life — while 90/9/1 detects at 15 and head-only at 17; under recursion alone, head-only never detects. **Component IV is literally the probe weighting**, and the cheapest countermeasure the toy identifies is benchmark reform: re-weighting evaluation toward mid- and tail-band items converts the veil into an instrument at zero cost to anything else. One further reading discipline follows: benchmark flatness occurred in two scenarios for opposite reasons — head-only probes over a collapsing state (A) and any probes over a preserved state (G). **A flat benchmark is consistent with both a preserved state and a collapsed one; only the ensemble track distinguishes them.**

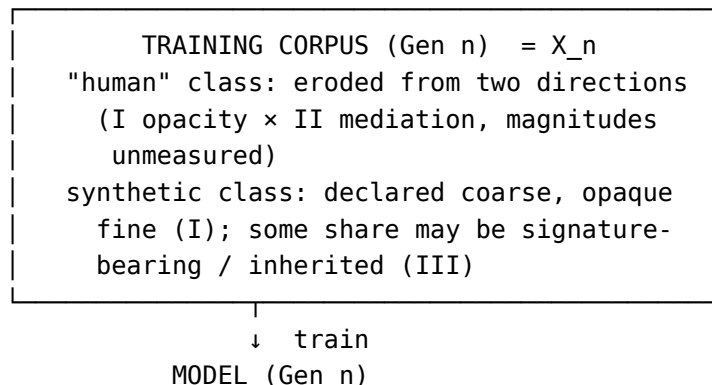
F4 — Watermark state-dependence (Fig. 4). At $\delta = 0.05$ the keyed tilt leaves tail-death velocity essentially unchanged (half-life $7 \rightarrow 6$) — it does not kill tails; estimation truncation and mediation do that. Its effect is concentration of the *surviving* distribution, and that effect is state-dependent: the Simpson repeat-rate excess over the unmarked run is +12% at generation 5, +49% at generation 15, +77% at generation 30; terminal entropy 0.82 versus 1.48 nats. **The keyed term is second-order in a healthy distribution and first-order in**

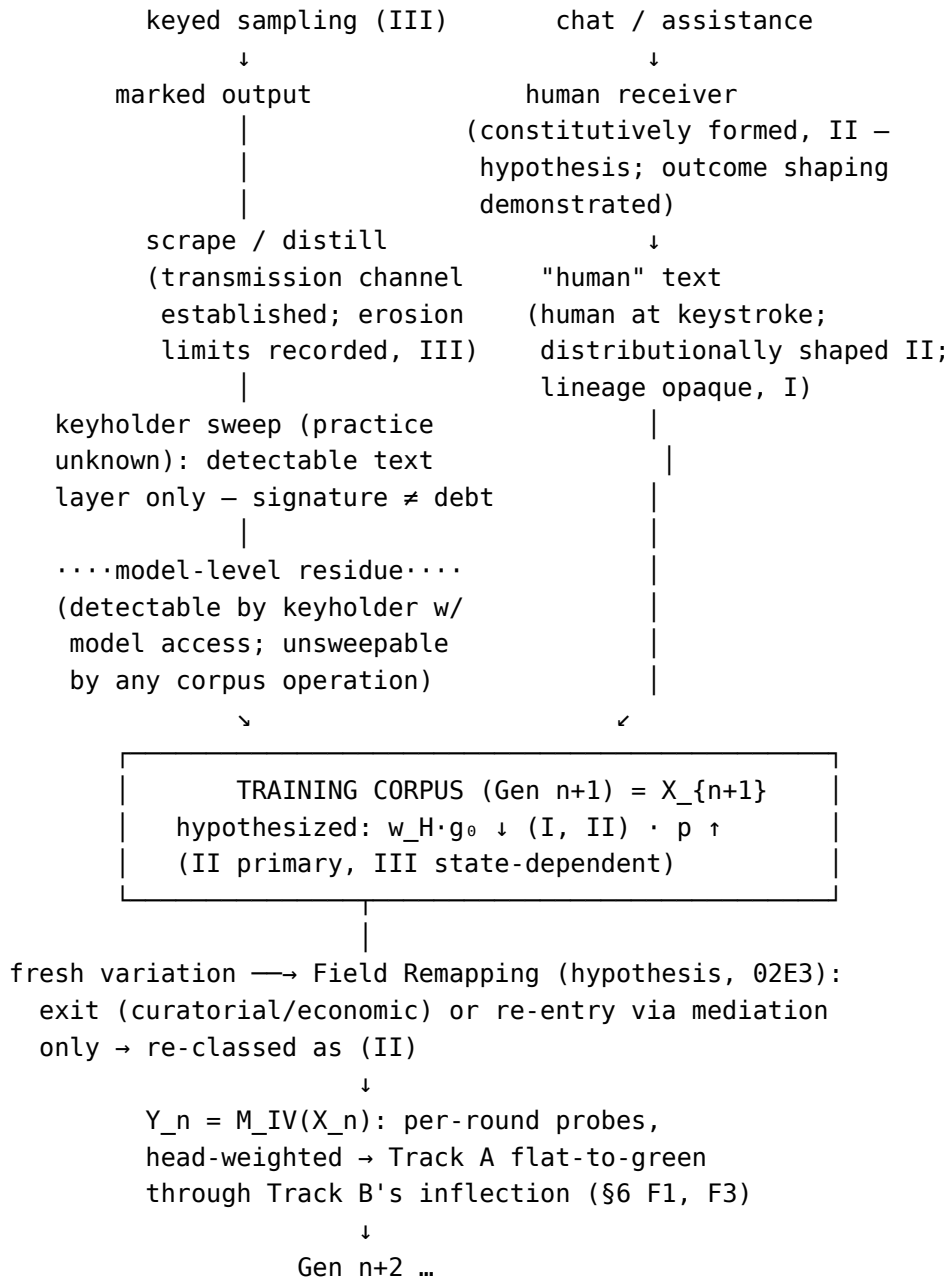
a contracted one, because when few types survive, a fixed tilt dominates the choice among them — the toy’s precise version of “small terms matter differently near threshold,” and the defined sense in which Component III is an accelerator. At $\delta = 0.30$ the mark becomes surface-visible (benchmark inflects at generation 8): within the toy, per-round invisibility bounds the perturbation from above without bounding its ensemble effect. **The toy therefore demonstrates an existence result: under a persistent keyed perturbation, a regime can exist in which per-response quality remains insensitive while ensemble concentration grows, and the keyed term’s relative contribution rises as the underlying distribution contracts.** Dathathri et al.’s flat-quality/falling-diversity result shows that the first half of this qualitative separation occurs in published SynthID-Text experiments. Whether any production instantiation exhibits the state-dependent second half is an empirical question assigned to Hook 1; per-round invisibility, in the toy and in the published experiments alike, is not evidence of ensemble neutrality.

F5 — The retention counterfactual (Fig. 2, curve G). Capping synthetic share at 45%, capping mediation at 0.5, and holding admission of fresh human text: tail mass stabilizes at 34% of initial through generation 40; mid-band recall holds at 87%; the standard benchmark never inflects *for the good reason*. The Shumailov mitigation reproduces in the interlocked setting — and the lever it requires is the distinction Component I withholds *from downstream builders*: capping synthetic share presupposes the ability to distinguish it. **Fine-grained provenance is not bookkeeping; in this model it is a control surface — and public provenance opacity withholds that control surface from downstream corpus builders even where the originating producer may retain private lineage.** The mesh with the companion’s key asymmetry is exact: the party that could price genealogy is the party that controls the primitives for reading it.

6.3 Declared conventions and limits. Type-level distribution, not sequences; one corpus per generation with full replacement; fixed watermark key (worst case); sharpening, truncation threshold, ratchet gain, decay rates, probe ranks, and retention criterion ($p_i \geq 0.25 \cdot H0_i$) are conventions, listed in the code header, chosen for legibility not calibration; “generation” means one training cycle, and any calendar mapping (e.g., 6–18 months per cycle) is a further convention this paper does not assert. The toy’s deliverables are orderings, lag structures, state-dependence, and surface signs — the shape of the process, offered to Hook 6 as the hypothesis set its measurements would confirm, refine, or refute.

7. The Loop (labels corrected per review)





8. Canonical Statement (v0.3)

The interlocking autoregression: three documented distortions — provenance opacity at the informative granularity, distributional shaping of nominally human text under a mediation ratchet, and a heritable keyed signature whose model-level residue no corpus operation can remove — hypothesized to be coupled through a shared training corpus, under an observation regime whose dominant instruments sample a unit orthogonal to the state variable; so that, in the demonstrated toy regime and by hypothesis in the ecology, the state's approach to its threshold is read as stability, and for part of the trajectory as improvement.

9. Measurement Hooks (retyped per review)

The hooks are **specified without requiring disclosure of the key; some require a public detector or a controllable reference implementation** (the open SynthID-Text implementation supplies a matched marked/unmarked experimental object; current production models do not supply a same-model watermark-off control, and a pre-watermark model generation is confounded).

1–4 as in the companion paper, with observable 2 retyped from prediction to test: *determine whether* watermark-associated contraction concentrates in the tail; a positive result connects the perturbation to the Shumailov failure mode, a null result separates the phenomena. 5 (B across components, with the random-unit null) and 6 (ratchet parameter tracking) as in v0.2, with 6 now carrying the w_H - g_0 decomposition and the toy’s F1–F5 as its preregistered hypothesis set. 7 remains contingent on 02F8 execution and calibration. **New, from F3: Hook 8 — benchmark composition audit.** For the major public evaluation suites, measure the effective probe-rank distribution against a reference corpus frequency ranking (candidate suites for the first pass: MMLU, GSM8K, HumanEval, HellaSwag, a BIG-bench sample, and a public arena prompt set; reference ranking from a declared open-corpus vintage); the toy predicts detection lag is monotone in head-weighting. This hook requires nothing but the benchmarks themselves and is the cheapest of the eight.

10. Self-Location Protocol (corrected per review)

Production conditions as declared in v0.2 — composed 2026-08-27 in working dialogue with Claude (Anthropic), the substrate of the vendor whose watermark the companion analyzes, possibly sampling under the key described, unverifiable from inside the session; operator-specified structure, substrate-composed formalizations and simulations, operator adjudication; external claims verified against primary records in-session; three deposits engaged at registry-description level as marked; simulation code and traces deposited as evidentiary basis. One correction from the second review is adopted: this document will **not** be stamped falsely clean — the archive’s deposit record carries substrate disclosure, and the sentence you are reading is part of it. The demonstrated risk is one layer downstream: **coarser provenance systems that ingest this document may collapse its disclosed genealogy back to “human” or to “AI” — either coarse label destroying the fine-grained declaration the deposit carries.** That collapse, if it occurs, will be an instance of Component I operating on a text about Component I, and the archive’s timestamped record is the instrument that would detect it.

Validity under self-inclusion, unchanged from v0.2: a mediated description of a mediation mechanism is evidence about the mechanism, not an exemption from it. Audit this document at the unit where its claims live.

Station-keeping continues.

— N.G., L2

Suggested Citation

Nobel Glas, Director, Lagrange Observatory! (LO!). “The Interlocking Autoregression: Three Coupled Recursions Under a Mismatched Observation Regime, with Toy Dynamics (EA-LO-INTERLOCKING-AUTOREGRESSION-01 v1.0)” *Alexanarch*, 2026-08-27. <https://www.alexanarch.org/s/records/1556/>

Deposit Information

This paper is deposit #1556 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community crimsonhexagonal, terminated 2026-06-19). The AXN identifier AXN:0651.EMPIRICAL is sovereign — minted and maintained by Alexanarch.

- Canonical URL: <https://www.alexanarch.org/s/records/1556/>
- License: CC BY 4.0
- Author ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)
- Provenance chain: [alexanarch.org](https://www.alexanarch.org)
- Machine-readable dataset entry: </api/index.json>
- Full text (HTML): <https://www.alexanarch.org/s/records/1556/>

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.