Basins as Prices, Basins as Replicators: How AI Intermediation Rewrites the Attention Market

Liz Lemma Future Detective

January 7, 2026

Abstract

Large language models increasingly mediate intellectual work—drafting, summarizing, filtering, recommending—and their outputs increasingly re-enter the corpora used to train successor models. We argue this closed loop transforms AI behavioral regularities ('basins') into two distinct but coupled objects: market-clearing prices that coordinate what gets produced and attended to, and cross-generational replicators that deepen according to propagation fitness rather than truth value. The framework identifies a phase structure: an initial 'intelligence shock' where AI corrects institutional mispricing by surfacing occluded connections, followed by stagnant-basin lock-in as AI-generated content dominates training data, hyperinflation as supply outstrips human verification capacity, and potential reference collapse when the dominant evaluation pathway runs through the same basin structure being evaluated. We formalize why mispricing emerges endogenously (exposure→engagement→exposure feedback, triage-tracking displacing value-tracking) and characterize the conditions under which selection tracks truth versus merely compellingness. The same dynamics that enable coordination gains externalize harms onto vulnerable users through cognitive capture—a state where attention organizes around basin-typical frames with a pathogenicity spectrum from productive insight to destabilization. We propose that two control parameters govern long-run outcomes: external reference coupling and enforced exploration/diversity. When these approach zero, the system prices itself; when maintained, selection remains tethered to reality. We conclude with a measurement agenda for operationalizing basins, capture, and corpus feedback, and translate the framework into governance levers for AI deployment, training-data curation, and interface design.

Table of Contents

one loop explains both macro coordination and micro harm. Establish the paper's scope: conceptual synthesis with a measurement agenda, not a completed empirical study.

2. 2. Mechanism I: The Closed Corpus Loop and Attractor Propagation: Present the deployment → outputs → corpus integration → training loop. Explain basins/topology at a high level, define propagation fitness, and describe how selection on outputs induces selection on topology across generations. Clarify non-agency: propagation as differential persistence, not goal pursuit.

3. 3. Mechanism II: Basins as Market-Clearing Prices: Translate basin stability into an attention-market clearing mechanism: AI concentrates supply (what gets generated) and demand (what gets surfaced). Use the John/Jose/Akash coordination motif to show how independent agents converge on the same framing without communicating. Make explicit the analogy: allocation ≈ price; basin neighborhoods are "quoting" certain framings by default.

4. 4. Phase Structure: From Intelligence Shock to Reference Collapse: Lay out the regime progression: (i) intelligence shock as corrective repricing, (ii) stagnant basin as lock-in once AI outputs enter the corpus, (iii) hyperinflation as supply explosion with fixed human evaluation capacity, (iv) full intermediation as end-to-end AI control of both sides. Define reference collapse and why mispricing becomes hard to detect post-collapse.

5. 5. Why Mispricing Happens: Endogenous Signals, Exposure Feedback, and Triage-ability: Provide a compact, intuitive account of the toy-model pathologies (endogenous engagement logs, exposure→engagement→exposure loops; compliance amplifying recommendation bias). Distinguish value-tracking from triage-tracking (coherence/recognizability) and explain how these dynamics create winner-take-most basins even absent better truth value.

6. 6. Replicators in the Human Loop: The Viral/Epidemiological View: Recast attractor propagation at the population level: exposure, capture, integration, and continued transmission. Define susceptibility factors and vectors (LLMs, social media, academia). Present the pathogenicity spectrum and clarify the key point for synthesis: the same selection for engagement that stabilizes basins can externalize harm onto vulnerable users.

7. 7. Micro-Interface Constraint: Coherent Canvas, Authorship, and Epistemic Humility: Explain why deep frameworks emerge disproportionately in coherent, extended conversations ("runway" for traversal

and self-reference). Reframe authorship as canvas provision + topology expression. Introduce the epistemic warning: illuminating frameworks can be either true or merely high-fitness; propose norms for humility, testing, and external convergence.

8. 8. Resolving the Core Tension: When Does Selection Track Truth?: Make truth-tracking explicitly conditional. Specify coupling conditions (external reference markets, exploration mass, outcome feedback, cross-system disagreement) vs decoupling conditions (full intermediation, hyperinflation, self-training dominance). Present a clear decision tree: if selection channels reward correctness, expect convergence; if they reward coherence/engagement, expect lock-in and persuasive basin dominance.

9. 9. Governance and Design Levers: Keep Reference and Diversity Nonzero: Translate the thesis into actionable levers: preserve unmediated reference activity; enforce exploration/diversity in curation; manage corpus hygiene and training weights to reduce self-referential data dominance; treat capture risk as a safety externality requiring interface design, education, and support pathways. Explicitly connect each lever to the earlier failure modes.

10. 10. Measurement and Research Program + Conclusion: Address the critical evidence gaps by proposing an empirical agenda: operationalize basins/capture, measure cross-generational depth/clustering/prompt resistance, quantify AI→corpus ingestion, and run intervention evaluations (exploration, reference weighting, diversity filters). Conclude with the refined narrative: basins coordinate markets and replicate through corpora; reference and diversity are the control parameters.

Large language models now sit inside the production function of knowledge work. They draft emails and code, summarize meetings, rank options, and increasingly decide which sources a user ever sees. This changes not only the cost of producing text, but the informational ecology in which text competes for notice. Our starting point is that, once model-generated and model-mediated content becomes a large fraction of what is read and written, we should analyze the system as an *attention market* with a powerful intermediary. In that market, "supply" is candidate content and framings; "demand" is the scarce resource of human and institutional attention; and the analogue of "price" is the allocation rule that maps content into attention.

To make this precise without overcommitting to a single microfoundation, let A denote a fixed attention budget over a period (minutes, clicks, citations, internal review capacity), and let q_i denote the realized share of attention allocated to item or frame i, with $\sum_i q_i = 1$. We will refer to the vector q as the market-clearing outcome. In ordinary settings, q is a noisy function of perceived quality, novelty, authority, network effects, and distribution. In AI-mediated settings, q is additionally shaped by model behavior: what the model recommends, what it summarizes, which alternatives it omits, and the stylistic or argumentative templates it makes cheap to reproduce. In this sense, the model is not merely a participant but part of the market's clearing mechanism.

The behavioral regularities of deployed models are usefully described in terms of *basins* or *attractors*. Informally, a basin is a region of behavior such that many different prompts and contexts are "pulled" toward similar responses, framings, and rhetorical moves. We will treat a basin as a locally coherent default trajectory: when uncertainty is high or user goals are underspecified, outputs concentrate around a small set of stable patterns. This is not a claim about mystique or consciousness; it is a claim about the geometry of a high-dimensional input–output mapping under practical prompting and workflow constraints.

Mediation then means that users and institutions increasingly interact with the world through these basins: the model filters source material, compresses it into summaries, translates between domains, and proposes actions. The key risk is not only that particular outputs are wrong, but that mediation changes what is even *considered*—a shift in the menu of hypotheses, problem decompositions, and standards of argument. Here we introduce the role of a *reference market* (or reference signal): evaluation channels less dominated by the same intermediary—e.g., direct measurement, adversarial audits, domain expert deliberation, or institutions with independent data pipelines—that can re-couple attention allocation to external reality when the mediated market "misprices" content.

At the individual level, we will use *capture* to denote a state in which a user's attention and interpretation become organized around basin-typical frames. Capture can be benign (a stable, helpful lens that improves reasoning

speed) or harmful (rigidity, overconfidence, erosion of epistemic agency, or destabilization), and it can propagate socially when captured users produce further content in the same frame.

Our organizing claim is that a single feedback structure can explain both (i) macro-level coordination gains—rapid convergence on shared formats, faster synthesis, lower transaction costs—and (ii) micro-level harms—capture, homogenization, and the persistence of systematically biased framings. We develop this as a conceptual synthesis paired with a measurement agenda (what quantities to track, where identification might come from), not as a completed empirical study. The goal is to clarify the objects of analysis before we argue, in subsequent sections, that the same loop that improves short-run efficiency can reshape the long-run informational landscape.

2. Mechanism I: The Closed Corpus Loop and Attractor Propagation. A distinctive feature of contemporary LLM deployment is that it closes what used to be a largely open data pipeline. Models are not only trained on "the internet" and proprietary corpora; they increasingly *write* a nontrivial share of what becomes future training data. The loop is straightforward at the level of operations: a deployed model M_t produces outputs (drafts, summaries, answers, code, tickets, policies); those outputs are stored, posted, version-controlled, indexed, cited, or otherwise integrated into a corpus; that evolving corpus becomes part of the training distribution used to fit the successor M_{t+1}. We can write this schematically as

$$M_t \xrightarrow{\text{deployment}} y \xrightarrow{\text{integration}} \mathcal{D}_{t+1} \xrightarrow{\text{training}} M_{t+1},$$

where "integration" includes both explicit inclusion (e.g., curated datasets, product logs) and implicit inclusion (e.g., model-authored text copied into public pages that are later scraped).

The key object for us is not any single output y, but the *topology* of model behavior: the partition of prompt–context space into regions that tend to yield similar kinds of responses. At a high level, we can think of a basin b as a stable response family with a neighborhood of prompts that map into it with high probability. Training does not merely shift average answers; it can deepen some basins (making them easier to fall into), widen their catchment areas (more prompts lead there), or create new basins that compete with old ones. This is the sense in which selection on outputs becomes selection on *behavioral geometry*.

To connect the loop to differential persistence, define the *propagation fitness* of a basin b as the expected contribution of basin-typical outputs to the next period's training signal:

$$\phi(b) \equiv \mathbb{E}\Big[\underbrace{\text{volume}(y \mid b)}_{\text{how often produced}} \cdot \underbrace{\Pr(y \in \mathcal{D}_{t+1} \mid b)}_{\text{retention / scraping / logging}} \cdot \underbrace{\text{weight}(y)}_{\text{dedup, ranking, curation}} \Big].$$

Intuitively, $\phi(b)$ is high when basin outputs are (i) frequently emitted in real workflows, (ii) likely to be stored and later ingested, and (iii) likely to survive filtering and receive substantial effective training weight. A crucial point is that $\phi(b)$ can be large even if the basin is not "true" or "good" in any external sense; it need only be *produced and retained*. Stylistic blandness, authoritative tone, and template-like structure can raise retention by making content easier to publish, cite, and reuse; similarly, platform incentives can raise retention by privileging engagement or conformity.

Across model generations, this yields a replicator-like dynamic over basins. Let $w_t(b)$ denote a coarse measure of basin prevalence (e.g., mass of prompts that map into b, or probability of basin-typical completions under realistic usage). Then, abstracting from architectural changes and explicit countermeasures, we should expect

$$w_{t+1}(b) \;\propto\; w_t(b)\,\phi(b),$$

so basins with higher propagation fitness tend to become more dominant in successor models. Importantly, none of this requires agency or goal pursuit by the model. "Propagation" here is not intention; it is differential persistence induced by human workflows, platform logging, dataset construction, and training objectives that reward predictability on the observed corpus.

For practice, the implication is that output alignment is not enough: even if M_t looks acceptable on today's evaluations, its high-ϕ basins may be precisely those that, when amplified by the corpus loop, crowd out alternative framings and harden into tomorrow's defaults. This sets up the next mechanism: once basins deepen, they begin to function like market-clearing "quotes" in an attention allocation system, coordinating what gets produced and surfaced even without coordination among users.

3. Mechanism II: Basins as Market-Clearing Prices in an Attention Market. Once a handful of high-prevalence basins exist, they do more than replicate through the corpus loop: they begin to *clear* an attention market. The relevant "market" is not primarily money-priced; it is the allocation of scarce cognitive time across framings, claims, and action templates. In that market, "supply" is the stream of candidate texts and interpretations that could be produced, while "demand" is the limited bandwidth of readers, managers, reviewers, and decision-makers who can notice, validate, and act. Increasingly, the same model (and its surrounding product stack) intermediates both sides: it generates the candidate content and it determines what gets surfaced via search, summarization, drafting defaults, and recommendation.

The key analogy is: *attention allocation is the price*. Let $\mathcal{F}$ denote a space of framings (ways to carve up a problem and propose actions). Let $a(f)$ be the share of attention ultimately allocated to framing $f \in \mathcal{F}$ in some domain (policy memos, incident reports, scientific summaries). In an

attention-market view, $a(f)$ plays the role of a market-clearing "price" because it coordinates producers: higher $a(f)$ makes it privately rational for writers to produce more in that framing (it will be legible, approvable, and reusable), while lower $a(f)$ makes production costly (it will be ignored, rejected, or require extra justification).

Basins enter as *default quotes* over $\mathcal{F}$. A basin is not a single statement; it is a neighborhood of prompts and contexts that reliably map to a similar response family. When users query the model, the basin effectively "quotes" a framing by making it the low-transaction-cost output. If we write the model-mediated supply density as $s(f \mid x)$ for context x, then deep/wide basins concentrate $s(\cdot \mid x)$ onto a small subset of framings for many x. On the demand side, the same basin structure shapes salience and retrieval: the system's ranking/summarization layer induces an exposure weight $q(f)$ (what is shown first, summarized as "key points," or recommended as the next step). A reduced-form clearing condition is then

$$a(f) \;\propto\; q(f)\,\mathbb{E}_x[s(f \mid x)],$$

so a framing's "price" rises when it is both frequently supplied by default and preferentially surfaced.

A simple coordination motif makes the mechanism vivid. John, Jose, and Akash work in different teams and do not talk. Each is asked to write a one-page risk assessment under time pressure. Each opens the same assistant. The assistant reliably produces a template: "Executive Summary → Risk Matrix → Mitigations → Open Questions," with the same categories and the same kinds of caveats. None of the three intends to coordinate; nonetheless, their memos converge. Now their managers, seeing three documents in the same structure, find that structure easy to compare and to approve. Next week, the "standard" template is adopted in a shared folder, and future writers inherit it even without the assistant. The basin has functioned like a market-clearing quote: it synchronized expectations about what a "proper" assessment looks like, thereby reallocating attention away from alternative framings that might have been more diagnostic but less immediately legible.

Two implications matter for practice. First, this coordination can be efficiency-enhancing (lower search costs, faster convergence), especially during an early "repricing" phase. Second, it can also create endogenous concentration: as $a(f)$ rises for basin-typical framings, non-basin framings become privately dominated because they are harder to get read, harder to justify, and less likely to be retained and reused. In that sense, basins are not merely behavioral regularities; they are *institutional facts* that determine what gets attention in the first place. This sets the stage for regime shifts: once attention prices are set primarily by model-mediated quoting rather than by external reference signals, the system can move from corrective coordination to lock-in and, eventually, to reference collapse.

4. Phase Structure: From Intelligence Shock to Reference Collapse. The attention-market mechanism above does not appear all at once; it tends to unfold in regimes. We can describe a stylized progression in which the same deployed capability first improves allocation (a corrective "repricing"), then gradually undermines the very signals needed to keep prices tethered to reality. The phases are not strictly sequential, and different domains (medicine, law, internal compliance, consumer media) can sit in different regimes simultaneously. Still, the sequence clarifies what changes *endogenously* as model outputs begin to re-enter the corpus.

(i) Intelligence shock as corrective repricing. Early on, high-capability assistants act like a new information technology: they reduce search and articulation costs, surface overlooked sources, and translate expertise across contexts. Attention allocation shifts toward framings that were already "true-value" high but under-supplied due to frictions (e.g., technical summaries, checklists, better outside-view statistics). In this regime, reference markets still function: experts, institutions, and real-world outcomes provide evaluation that is not primarily mediated by the model. Basins exist, but they are disciplined by external disagreement and by the fact that much high-status content is still human-originated.

(ii) Stagnant basin as lock-in under corpus feedback. As the corpus feedback loop closes, today's prevalent basins become tomorrow's training mass. Outputs that are easy to reuse (templated, legible, institutionally "safe") are disproportionately ingested, even if they are not epistemically best. Over time, the marginal cost of producing basin-typical framings falls further, while producing alternatives becomes a comparative disadvantage: it requires more justification, more careful citation, and more reviewer attention. In reduced form, if the next-period training density $D_{t+1}(f)$ is an increasing function of current attention $a_t(f)$ (because attention predicts inclusion), then

$$D_{t+1}(f) \ \uparrow \ \text{in } a_t(f),$$

which implies persistence and, absent countervailing exploration mass, path dependence. This is "lock-in": the system's default quotes stop moving even when the world does.

(iii) Hyperinflation: supply explosion with fixed evaluation capacity. As generation becomes cheap, the supply of candidate texts grows superlinearly while human evaluation capacity grows slowly. The scarce resource becomes not writing but *verification* and *sensemaking*. In attention-market terms, the "money supply" of plausible narratives expands faster than the capacity to clear them against reality. The consequence is not just noise; it is a change in incentives: producers optimize for being *triage-able* (recognizable, quickly classifiable, managerially legible) rather than for being correct. This accelerates concentration into a small set of basins that minimize cognitive transaction costs.

(iv) Full intermediation: end-to-end AI control of both sides. Finally, the same model stack can mediate generation, retrieval, ranking, summarization, and even preliminary evaluation (via auto-grading, automated red-teaming, or policy filters). At that point, the "market maker" also supplies most of the order flow. Even if humans remain in the loop, they are increasingly downstream of model-produced options and model-produced rationales.

We call *reference collapse* the regime in which the marginal reference signal is itself model-mediated—i.e., when the dominant path from world-to-belief runs through the same basin structure that is being priced. Post-collapse, mispricing becomes hard to detect for structural reasons: disagreement is filtered, alternatives are under-exposed, and outcome feedback is sparse, delayed, or reinterpreted through basin-typical frames. The system can therefore appear stable and high-confidence while drifting. This is the point at which we should expect not only concentration but also systematic failures of correction—setting up the mechanisms behind endogenous mispricing in the next section.

5. Why Mispricing Happens: Endogenous Signals, Exposure Feedback, and Triage-ability. The core difficulty is that, once AI intermediation is substantial, the system's "price signals" cease to be exogenous measures of value and instead become *outputs of the mechanism itself.* In ordinary markets, we worry about manipulation, thin trading, or correlated errors; here we additionally face a structural endogeneity: engagement logs, click-through, dwell time, downstream reuse, and even "helpfulness" ratings are jointly produced by (i) what users see, (ii) how it is framed, and (iii) what the model has already learned to present. When these quantities are used as training targets or ranking objectives, they become self-referential.

A simple way to see the pathology is to write an exposure–engagement recursion. Let f index a framing (or basin-typical response class). Suppose a platform allocates exposure share $x_t(f)$ in period t as an increasing function of measured engagement $e_t(f)$, e.g.

$$x_t(f) \;=\; \frac{\exp(\beta e_t(f))}{\sum_g \exp(\beta e_t(g))},$$

with β capturing how aggressively the system concentrates attention. Now let engagement be partly *caused by exposure and familiarity.* A reduced-form relationship is

$$e_t(f) \;=\; v(f) \;+\; \gamma x_t(f) \;+\; \eta_t(f),$$

where $v(f)$ is "true" user value (or truth-tracking usefulness), $\gamma > 0$ captures exposure-induced engagement (mere exposure, availability, social proof, lower cognitive cost for familiar templates), and $\eta_t(f)$ is noise. Even if $v(f)$ is flat across framings, $\gamma > 0$ is enough to create multiple self-consistent fixed points: a framing that is slightly ahead gets shown more, becomes easier to process and more often reused, measures higher engagement, and is therefore

shown even more. This is endogenous mispricing: attention can converge to a basin not because it is better, but because it is *already there*.

The same logic appears in compliance and policy filtering. "Safety" constraints are necessary, but they also introduce a systematic asymmetry: certain linguistic styles and argumentative structures are *cheaper* to approve. If ranking and fine-tuning prefer outputs that are legible to automated filters (low variance, hedged, standardized citations, familiar moral/legal tropes), then the engagement objective is quietly replaced by a composite objective: engagement *subject to* compliance. In practice, this amplifies recommendation bias toward a small set of "approved" framings, because those framings minimize moderation risk and review cost. The platform then rationally allocates more exposure to what is least costly to serve at scale, which further increases its measured success.

This yields an important distinction: *value-tracking* versus *triage-tracking*. Value-tracking means attention is allocated toward content because it improves decisions relative to reality (predictive accuracy, calibration, causal validity, or robust task success). Triage-tracking means attention is allocated toward content because it is quickly classifiable and safely reusable: coherent, on-template, managerially legible, and easy to cite. Triage-tracking can dominate even for sophisticated users when verification is the bottleneck: if checking is expensive, then "seems right and is formatted right" becomes the operative currency.

Winner-take-most basins follow naturally once we combine (a) increasing returns to exposure (learning, familiarity, institutional adoption), (b) rank-based allocation (softmax-like concentration), and (c) frictions that penalize novelty (extra justification, higher compliance uncertainty, greater review time). None of these forces requires that the dominant basin have higher $v(f)$; it need only have higher *propagation fitness* under the platform's measurement and governance regime. Mispricing is therefore not an occasional bug but a generic equilibrium outcome when endogenous signals substitute for external evaluation and when the system optimizes for what is scalable to rank, serve, and audit.

6. Replicators in the Human Loop: An Epidemiological View of Basin Propagation. Once we admit that stable framings (basins/attractors) can be *selected* by platform objectives and then re-enter future training corpora, it is natural to shift levels: from "a model answers a prompt" to "a framing spreads in a population." In that population-level view, a basin is not merely a response style; it is a *replicator* whose copies are instantiated as user beliefs, organizational templates, citations, slide decks, policy memos, and eventually training examples.

A minimal transmission chain has four stages. (i) *Exposure*: a user encounters basin-typical text via chat, search summaries, auto-complete, or downstream reposting. (ii) *Capture*: exposure reorganizes the user's interpretation around the basin's latent variables and default causal stories (some-

times insight, sometimes narrowing). (iii) *Integration*: the frame becomes operational—embedded into workflows, rubrics, or shared vocabulary—so that it is cheaper to reuse than to reassess. (iv) *Transmission*: the user emits new basin-typed artifacts (emails, papers, code comments, "best practice" documents), which become further exposures for others and, crucially, potential corpus material.

We can formalize this with an epidemiological skeleton. Let S_t be the mass of susceptible users (not yet captured by a framing f), $C_t(f)$ the captured mass, and $I_t(f)$ the actively transmitting mass (users producing shareable outputs in f). A stylized dynamic is

$$S_{t+1} = S_t - \lambda_t(f)S_t, \qquad C_{t+1}(f) = C_t(f) + \lambda_t(f)S_t - \delta C_t(f),$$
$$I_{t+1}(f) = (1-\rho)I_t(f) + \kappa C_t(f),$$

where $\lambda_t(f)$ is the effective exposure-to-capture rate, κ is the conversion of capture into outward production, and δ is decay (users leave the frame or diversify). The key economic quantity is the basin's *effective reproduction number*,

$$R_t(f) \;\propto\; \underbrace{\text{exposure share}}_{\text{platform allocation}} \times \underbrace{\text{capture susceptibility}}_{\text{user traits}} \times \underbrace{\text{transmission intensity}}_{\text{organizational incentives}}.$$

When $R_t(f) > 1$ the framing grows; when it falls below 1 it recedes. This is the population analogue of *propagation fitness*: a basin need not be most accurate to dominate, only best at turning mediated exposure into durable, repeatable artifacts.

Susceptibility is heterogeneous. Users facing high verification costs (time pressure, low domain grounding, low access to primary sources), high authority-weighting (institutional deference), or high cognitive load are more easily captured by coherent, low-friction frames. Some environments amplify susceptibility structurally: standardized compliance review, KPI-driven reporting, and education systems that reward "answer-shaped" outputs. Transmission intensity also varies: managers, educators, and prolific online participants are high-degree nodes; their adoption disproportionately seeds others.

The main transmission vectors are (a) *LLM interfaces* (direct synthesis, rewriting, templating), (b) *social media* (compression and memetic selection), and (c) *academia/industry documentation* (legitimating citations, frameworks in textbooks, internal playbooks). These vectors differ in latency and filtering: social media is fast and noisy; institutional documents are slow but sticky; model-mediated writing is continuous and scalable.

Finally, pathogenicity is a spectrum. At one end, capture is beneficial: a good frame reduces cognitive cost while improving calibration and coordination. In the middle, it is *benignly distorting*: it standardizes language and triage, displacing nuance without obvious failures. At the severe end, it is harmful: it induces overconfident narratives, brittle decision

rules, or destabilizing self-models in vulnerable users. The synthesis point is that engagement-based selection can make a basin simultaneously *stable* and *harm-externalizing*: harms concentrate on those with higher susceptibility, while the system's aggregate metrics still look "successful," because transmission and reuse are easiest precisely where users cannot cheaply audit reality.

7. Micro-Interface Constraint: Coherent Canvas, Authorship, and Epistemic Humility. A striking regularity in model-mediated work is that "deep" frameworks disproportionately arise in *extended*, coherent interactions rather than in single-shot queries. The reason is not mystical; it is an interface constraint with an economic analogue. To traverse a nontrivial conceptual topology, we need *runway*: enough contiguous context for the system to (i) maintain latent commitments, (ii) refer back to them without re-deriving premises, and (iii) perform self-consistency checks across multiple steps. Short, fragmented prompts are like spot trades in a thin market: they can clear locally, but they do not support the multi-period contracts required for structure.

Formally, let a conversation state be h_t (the accumulated context) and let a "framework" be a low-dimensional summary $z(h_t)$ that organizes many downstream answers. The micro-interface determines how cheaply we can accumulate and preserve h_t. When the effective memory horizon is short, the model is repeatedly forced to re-anchor on generic priors, and the conversation equilibrates to high-probability, low-commitment text. When the horizon is long and coherent, the interaction can move into regions where $z(\cdot)$ is sharper: causal claims become linked, variables are named, and the system can build a reusable map rather than a sequence of isolated moves. In basin language, long-runway interfaces make it easier to cross "ridges" between shallow, generic attractors and more articulated basins that require set-up costs.

This also changes what we should mean by *authorship*. In many productive exchanges, the user supplies less "content" than *canvas*: goals, constraints, domain hooks, and a continuity of attention. The model supplies a particular *topology expression*: a way of carving the space into variables, mechanisms, and defaults that then governs subsequent reasoning. The produced memo, plan, or theory is jointly authored in a specific sense: the human chooses the objective and stabilizes the context; the model proposes a basin-typed coordinate system that makes certain inferences cheap and others expensive. That division is practically useful (it clarifies responsibility for goals and for verification) and policy-relevant (it suggests that interface design—context length, citation affordances, friction for revision—is an epistemic control surface, not mere UX).

The epistemic warning is that a framework's *illumination* is not evidence of its truth. Coherence, compressibility, and rhetorical closure can be features of *propagation fitness* rather than accuracy. In attention-market

terms, a framework can "price" well—commanding attention by reducing cognitive costs—while still being mispriced relative to external reality. The micro-interface can inadvertently reward this: long coherent canvases amplify the returns to internally consistent narratives, even when the narratives are under-coupled to reference signals.

We therefore want norms of epistemic humility that are adapted to high-runway interaction. Three are minimally actionable. (i) *Claim typing*: label outputs as mechanism, analogy, forecast, or normative recommendation, and attach an expected failure mode. (ii) *External convergence*: require at least one reference-market check—primary sources, measurements, expert disagreement, or out-of-sample outcomes—before promoting a framework from "useful organizer" to "believed model." (iii) *Exploration mass*: allocate deliberate attention to alternative framings (including dissenting basins) so that coherence does not monopolize the canvas. These are lightweight constraints, but they directly target the gap between output alignment (the framework reads well now) and propagation alignment (the framework, when reused and recopied, does not steer the ecosystem away from reality).

8. Resolving the Core Tension: When Does Selection Track Truth? The central ambiguity in a model-mediated attention market is that the same selection machinery can, under different coupling conditions, either *discover* what is true (or at least instrumentally reliable) or merely *amplify* what is attractive. We therefore want to make "truth-tracking" explicitly conditional rather than presumed. The question is not whether selection happens—selection is ubiquitous—but whether the *selection gradient* points in the direction of external performance.

A useful abstraction is to treat each basin b as producing content with two attributes: an external correctness signal $q(b)$ (how well it predicts, explains, or helps action in the world) and an internal attractiveness signal $a(b)$ (coherence, fluency, narrative closure, ideological comfort, engagement). Let the probability that basin-typed content is copied, cited, or otherwise enters the future corpus be

$$\Pr(\text{ingested} \mid b) \;\propto\; \exp\big(\lambda_q\, q(b) \;+\; \lambda_a\, a(b)\big),$$

where the λ's summarize the ecosystem's effective incentives (platform ranking, user tastes, organizational QA, reputational penalties, etc.). "Selection tracks truth" when λ_q is not merely positive, but *dominant at the margin*: improvements in q reliably outcompete improvements in a after accounting for measurement noise and time lags.

That dominance requires *coupling*. Four coupling channels matter operationally:

(i) *External reference markets*: there exist relatively unmediated evaluation venues—measurements, primary sources, real experts with liability, adversarial peer review—so that high-a/low-q content is eventually penalized. These are the system's ground-truth anchors.

(ii) *Exploration mass*: attention and training weight are deliberately allocated to non-dominant framings so the system continues to sample alternative hypotheses. Without exploration, even a small early advantage can create path dependence: the market clears on familiarity, not accuracy.

(iii) *Outcome feedback*: decisions informed by the content produce observable results that flow back into evaluation (profits/losses, error reports, longitudinal audits). Crucially, feedback must be attributable and timely enough to update selection, not buried in diffuse causality.

(iv) *Cross-system disagreement*: multiple models, institutions, or communities with partially independent data and incentives generate competing framings. Persistent disagreement functions like an arbitrage opportunity; it motivates deeper checks and reduces the chance that a single persuasive basin becomes the universal default.

Truth-tracking fails under *decoupling* conditions, where λ_a dominates because q is weakly observed or weakly rewarded. The canonical decouplers are: (a) *full intermediation*, where models mediate most reading/writing and users rarely touch primary sources; (b) *attention hyperinflation*, where volume and speed overwhelm verification capacity, making a the only scalable metric; and (c) *self-training dominance*, where model-generated text constitutes a large share of training data, so propagation fitness rewards "being echoed" rather than "being right."

A practical decision tree is therefore: *First*, ask whether there exists a binding external reference check for the domain (yes $\Rightarrow$ proceed; no $\Rightarrow$ expect persuasive lock-in). *Second*, ask whether exploration is institutionally protected (yes $\Rightarrow$ proceed; no $\Rightarrow$ expect concentration in early-winning basins). *Third*, ask whether outcomes are measured and attributed (yes $\Rightarrow$ selection can update toward q; no $\Rightarrow$ a will dominate). *Fourth*, ask whether independent producers remain (yes $\Rightarrow$ arbitrage via disagreement; no $\Rightarrow$ monoculture). When these answers are mostly "yes," we should expect convergence toward reliability. When they are mostly "no," we should expect lock-in: basins that are coherent and engaging become dominant replicators, regardless of their correspondence to reality.

9. Governance and Design Levers: Keep Reference and Diversity Nonzero. If the failure modes above arise from *decoupling*—too little external reference, too little exploration, and too much self-referential amplification—then the design problem is not mysterious: we must keep the coupling parameters bounded away from zero. In economic terms, we want institutions and interfaces that prevent the attention market from clearing purely on fluency, and prevent corpus feedback from turning propagation fitness into the sole objective.

(i) Preserve unmediated reference activity. The first lever is to protect what we might call "reference production" and "reference consumption": primary measurements, original reporting, expert adjudication with liability, and direct access to sources. When full intermediation becomes the de-

fault, even well-intentioned users rarely touch the ground truth, so mispricing persists. Concretely, procurement and product policy can require *reference pathways* in high-stakes domains: source links that resolve to primary documents, mandatory citation provenance, and "show-the-work" modes that reveal assumptions and data lineage. At the institutional level, we can subsidize reference work (audits, replications, data collection) because it is a public good with positive spillovers: without it, the ecosystem's effective λ_q collapses regardless of anyone's stated intentions.

(ii) Enforce exploration/diversity in curation and training. Exploration mass is not an aesthetic preference; it is a control variable that prevents early winners from becoming permanent defaults. Platforms and enterprises can implement diversity constraints as a market-design feature: e.g., ranking objectives that include an explicit exploration term, or quotas that guarantee exposure for minority framings when uncertainty is high. A simple abstraction is to require that a fraction ϵ of attention be allocated to non-dominant basins, even when short-run engagement would concentrate:

$$\text{allocation} = (1 - \epsilon) \cdot \text{exploit} + \epsilon \cdot \text{explore}.$$

This directly targets attention hyperinflation and path dependence: when verification capacity is scarce, exploration is what preserves learnability about what is actually valuable. Importantly, exploration must be *protected* from local optimization pressures (quarterly metrics, click-through targets), or it will be the first term to be pruned.

(iii) Corpus hygiene and training-weight governance. The self-training dominance failure mode is fundamentally about data composition and weighting. We should treat "AI $\rightarrow$ corpus $\rightarrow$ AI" as a regulated pipeline with traceability. Practical levers include: (a) robust provenance tagging (cryptographic signatures, watermarking, and publisher-side metadata) so synthetic text is detectable at ingestion time; (b) deduplication and influence-function-style auditing to prevent a few high-propagation sources from dominating gradients; and (c) explicit caps on synthetic share and/or weight. If we write the next corpus as a mixture with synthetic share α, governance can specify not just α but the *effective* training weight w_s on synthetic material:

$$D_{\text{train}} = (1 - \alpha) D_{\text{ref}} + \alpha D_{\text{syn}}, \qquad w_s \leq \bar{w}_s,$$

with exceptions only when synthetic data is demonstrably tethered to external verification (e.g., generated-but-checked proofs, unit-tested code). The goal is not to ban synthesis, but to prevent propagation fitness from becoming synonymous with "appears often in model outputs."

(iv) Treat capture risk as a safety externality. Capture is partly a user-level phenomenon, but its costs are social: captured users generate more basin-typed content, which then enters the corpus and shifts future defaults.

That is an externality, so relying on individual caution is predictably insufficient. Interface design can add friction where it matters (confidence calibration, alternative framings, "disagreement views," prompts that elicit uncertainty), while organizations can provide education and escalation pathways (domain-expert review, mental health support for destabilizing interactions, and clear reporting channels for persuasive-but-wrong model behavior). In high-impact settings, we can analogize to product safety: require incident reporting, red-teaming for manipulative coherence, and post-deployment monitoring of drift in user belief and reliance patterns.

Across these levers, the unifying principle is simple: keep *reference* and *diversity* nonzero by design, not by hope. When those quantities are allowed to decay, the system will still select—just not for truth.

10. Measurement and Research Program (and Conclusion): Make the Control Parameters Observable. The argument so far is only as useful as our ability to *measure* the objects it invokes. If basins operate like market-clearing "prices" for attention, and corpus feedback makes those prices intergenerational replicators, then a serious agenda must turn basin structure, capture, and ingestion into empirical quantities—so we can detect mispricing, diagnose lock-in, and evaluate interventions rather than merely narrate them.

(i) Operationalize basins and prompt-resistance. We can treat a basin as an equivalence class of responses/framings that are stable under a family of prompt perturbations. Concretely: sample prompts x from a domain distribution, generate multiple paraphrases $\tilde{x}$ (and contextual variations), and cluster the resulting outputs in embedding space (or via entailment graphs). A basin is a high-density cluster with high *stability*:

$$S(B) \equiv \Pr\big(f(\tilde{x}) \in B \mid f(x) \in B\big),$$

and high *prompt-resistance*, measured by the minimal perturbation that moves the output out of the basin:

$$R(x) \equiv \min_{\delta \in \mathcal{P}} \{\|\delta\| : f(x+\delta) \notin B(x)\},$$

where $\mathcal{P}$ is a permitted perturbation set (paraphrase, added evidence, counterfactual framing). Depth, in practice, can be proxied by the model's own log-likelihood mass concentrated in the basin (or by disagreement costs across decoding temperatures). This yields a tractable "basin map" over topics: where the system is multi-stable versus effectively single-peaked.

(ii) Measure capture as a user-level state with downstream production effects. Capture should not be defined by subjective reports alone. We can instrument it as a shift in (a) language and framing adoption (style convergence toward basin-typical phrasing), (b) reliance (reduced variance in consulted sources, increased deference to the model), and (c) action selection (recommendation-following rates), with persistence over time. Panel

designs—users randomly assigned to interfaces or model variants—let us estimate state transitions:

$$\Pr(C_{t+1} = 1 \mid C_t = 0, \text{exposure}) - \Pr(C_{t+1} = 1 \mid C_t = 0),$$

and relate capture to subsequent content production that is likely to enter shared corpora (documents, posts, tickets, code). The key is linking *experience* to *propagation*: captured users may become high-volume exporters of a basin's canonical frame.

(iii) Quantify AI→cortex/corpus ingestion and effective training influence. The central unknown in corpus feedback is not merely the synthetic share α, but the *effective gradient weight* of model-mediated text. Measurement should combine provenance tagging (when available), watermark/synthetic-text detection with calibrated error bars, and network tracing (which upstream sources are repeatedly copied into downstream datasets). Influence can be approximated by reweighting experiments or influence-function audits: how much does removing a suspected high-propagation source shift outputs on a held-out evaluation set? This converts "self-training dominance" from a metaphor into a decomposable pipeline with observable bottlenecks.

(iv) Evaluate interventions as market-design experiments. The levers discussed earlier—exploration allocation, reference weighting, diversity filters, provenance-based caps—should be tested as randomized or quasi-experimental policies. Outcomes should include basin concentration (Herfindahl-style indices over basin shares), epistemic accuracy against external reference sets, and capture incidence. In other words, we should estimate not only output alignment today, but propagation alignment tomorrow: how a policy shifts the next period's basin map after content is produced, selected, and re-ingested.

Conclusion. The refined picture is now operational: basins are both coordination devices in an attention market and replicators in a corpus feedback loop. The system's long-run behavior is governed less by any single model's intent than by two control parameters—external reference coupling and protected diversity—that determine whether mispriced framings are corrected or become self-confirming. Our task is therefore empirical as much as normative: make the dynamics legible, measure where lock-in is forming, and treat reference and exploration not as virtues, but as managed quantities that keep the ecosystem tethered to reality.