The Geometry of Correction

Self-Reference, Reality-Anchoring, and the Design of Systems That Remain Answerable

Flyxion

July 2026

Preface

This book began as a shorter essay about a narrower subject: the persistence of impersonation accounts, coordinated bot activity, and fraudulent marketplaces on large social platforms, despite those platforms' evident technical sophistication in every domain that generates revenue. An earlier and substantially shorter version of that argument appeared under the title *The Hollow Network: Synthetic Sociality, Metric Misalignment, and the Politics of Legibility in Algorithmically Mediated Environments*. The present volume retains parts of that original analysis, particularly in Chapters 1 through 3, largely unchanged in argument if not in length, but expands the underlying framework considerably beyond its initial scope. The earlier essay is therefore best understood not as a separate work so much as a first articulation of the questions this book attempts to answer more generally.

What changed, in the course of extending that argument, was the discovery that the mechanism explaining platform behavior was not specific to platforms at all. The same structure — an evaluating system whose objective has stopped depending on some channel it does not control — recurs in the governance of large institutions, in the architecture of individual human cognition, and, recursively, in whatever is proposed to correct any of the above. Chapter 4 states that structure once, formally, as the distinction between self-referential and reality-anchored evaluation. Everything after it is an application of that one distinction, developed with enough care in three different registers — institutional, cognitive, architectural — that a reader should, by the end, be able to recognize the same pattern in domains this book never mentions.

A note on scope. This book resisted, deliberately, the temptation to demonstrate its central claim across every domain it might plausibly apply to — financial markets, scientific institutions, education, AI alignment more broadly than Chapter 15 treats it. Readers familiar with those domains will likely see the pattern there as well; that recognition is, in a sense, the intended outcome, and it did not require the book itself to make the argument seventeen more times. What follows

develops one distinction rigorously in three domains rather than gesturing at it across many. Whether that was the right editorial choice is, appropriately, not a question this book can answer for itself.

A note on the chapters drawing on the Free Energy Principle and Lacanian psychoanalysis (Chapters 10 and 11 particularly). These chapters answer a question the institutional argument alone cannot: not why platforms are built to capture attention, but why a mind that already understands it is being captured remains captured anyway. The governance argument developed in Parts II and IV does not depend on these chapters being correct, and a reader unpersuaded by their specific theoretical commitments will lose nothing in the chapters that do not rely on them. This is stated plainly, more than once, in the chapters themselves — not as a hedge, but because the book's overall argument is, I think, stronger for not requiring every reader to accept every one of its parts.

Introduction: How This Book Is Organized

This book is built around a single formal distinction, introduced once, in Chapter 4, and applied three times. The distinction is between *self-referential* evaluation, in which a system judges its own success entirely by signals it generates itself, and *reality-anchored* evaluation, in which that judgment remains coupled to some external channel the system does not fully control. A coupling coefficient, ρ, measures the strength of that connection. When $\rho \to 0$, a system's self-assessment has become effectively unfalsifiable by anything outside itself — not because the outside world has gone silent, but because the system has stopped listening to it. This last point, named in Chapter 4 as the Correction Suppression Principle, recurs throughout the book: pathological self-reference is far more often a matter of suppressed listening than of genuinely absent evidence.

Part I: The Geometry of Correction (Chapters 1–4). The book opens where its argument originated: with the specific, concrete puzzle of large social platforms that possess enormous technical sophistication yet remain saturated with crude identity fraud. Chapter 1 states the puzzle. Chapter 2 examines the rhetorical move by which platforms redescribe the puzzle as an unavoidable cost of open infrastructure. Chapter 3 gives the underlying mechanism a formal statement, showing precisely where authenticity fails to appear in a platform's optimization target. Chapter 4 generalizes all three chapters into a single distinction — self-reference versus reality-anchoring — general enough to apply to institutions, minds, and, recursively, to whatever is built to correct either one.

Part II: Institutional Self-Reference (Chapters 5–7). These chapters apply the framework to organizations. Chapter 5 argues against treating a platform as a single coherent agent, showing instead how self-reference emerges from the uncoordinated interaction of many locally rational subsystems. Chapter 6 examines a specific governance pattern — concentrated strategic authority paired with distributed operational responsibility — that makes this drift especially durable.

Chapter 7 traces where the resulting costs land, on individuals least equipped to bear them, and closes with the one respect in which the present condition is genuinely without precedent in the long history of anxiety about mediated communication.

Part III: Subjective Self-Reference (Chapters 9–13). These chapters shift from institutions to minds. Chapter 9 draws a necessary boundary: the presence of synthetic interaction is not itself the problem this book is concerned with; its covert, non-consensual imposition is. Chapter 10 describes the predictive architecture of human cognition and the specific vulnerability that architecture creates. Chapter 11 explains why that vulnerability persists even after a person consciously recognizes it. Chapter 12 describes the ongoing cognitive labor — waymaking — through which reality-contact is actively maintained, at every scale from a single moment of attention to a lifetime's sustained intellectual project. Chapter 13 shows how contemporary platform design, through the systematic elimination of friction, erodes exactly this capacity.

Part IV: Corrective Architectures (Chapters 14–17). The final part turns to remedy, constrained throughout by the recursion result established in Chapter 4: no corrective system is exempt from the framework it applies to everything else. Chapter 14 argues for distributed, sympoietic correction over centralized authority, on exactly these grounds. Chapter 15 explains why this question has acquired new urgency given the trajectory of contemporary AI development. Chapter 16 develops a concrete technical architecture — irreversible event histories combined with quadrangulated meaning-stabilization — through which the preceding two chapters' principles can actually be built. Chapter 17 closes the book by asking what institutional conditions would be required to adopt that architecture, and refuses, on the book's own terms, to claim that any institution proposed to enforce it is thereby exempt from the same scrutiny.

A note on reading order. The book is written to be read in sequence, and later chapters generally assume earlier ones. That said, a reader interested primarily in the institutional argument can read Chapters 1 through 7 and 14 through 17 as a complete, self-contained treatment, setting aside Chapters 9 through 13's cognitive register entirely. A reader interested primarily in the cognitive and psychoanalytic material can read Chapter 4 for the general framework and proceed directly to Chapters 9 through 13. Chapter 4 is, in this sense, the one chapter

every reading path requires.

Contents

Part I

The Geometry of Correction

Chapter 1

The Perceptual Contradiction

1.1. An Ordinary Login

Begin with an unremarkable evening. A person opens a social platform — not out of any particular purpose, simply the reflexive motion of a thumb toward an icon — and, in the space of a few minutes, encounters the following: a friend request from an account bearing the name and photographs of someone they already know, sent from a profile created three days ago; a video, watched to completion before they register the decision to watch it, in which a voice they do not recognize narrates footage they cannot place, assembled from fragments that resemble a dozen other videos they have half-watched before; a comment thread beneath a post about a local issue in which seven accounts, each with a handful of followers and a recently created join date, converge on an identical phrase within minutes of one another; an advertisement for a pet available for adoption, heartbreakingly specific in its details, that a friend later reports paying a deposit toward before the account and the pet both vanished.

None of this required the person to seek out anything unusual. It is not the residue of a bad neighborhood of the internet, deliberately sought or stumbled into. It is simply what using a mainstream platform now involves, distributed through the ordinary texture of an ordinary session, indistinguishable in its presentation from the messages of real friends and the posts of real communities that arrive alongside it.

This is the condition this book calls **synthetic sociality**: an environment in which automated, fabricated, or algorithmically amplified activity constitutes a large enough share of what a user encounters that genuine interaction can no longer be reliably distinguished from simulation by inspection alone. The term is offered descriptively, not as an accusation. It names a structural feature of an environ-

ment, not a moral verdict on the people who built it. Its purpose is to take an experience that is already familiar — the vague, low-grade suspicion that accompanies scrolling, the second look at a message before responding to it, the friend who says *wait, is that really you?* before answering a call — and make that experience available for the kind of systematic examination a private, individually-borne feeling of unease rarely receives.

1.2. The Puzzle Stated Plainly

Here is what makes the condition worth pausing on rather than simply enduring. The organizations that build and operate these platforms are not, by any reasonable account, unsophisticated. They are, by their own extensive public description, among the most advanced applied-machine-learning organizations that have ever existed. Their research divisions publish at the frontier of the field. Their infrastructure processes interaction data at a scale that has no precedent in human history. Their recommendation systems can, with well-documented precision, model an individual user's attention patterns closely enough to predict, action by action, what will hold that attention next — a capability advertisers pay enormous sums to access and one these companies do not undersell in their own communications about what artificial intelligence will do for content moderation, personalized discovery, and the deepening of human connection at scale.

Set that self-description beside Section 1.1. If an organization can model, with actionable precision, which sixty-second video will keep a specific sixteen-year-old's attention through four more minutes of scrolling, the claim that the same organization cannot reliably detect that an account created on Tuesday, using a photograph lifted from another account, is soliciting money from that photograph's actual owner's actual contacts, requires an explanation. The two problems are not obviously of comparable difficulty. Predicting attention from petabytes of behavioral signal is, by any conventional measure, the harder machine learning problem. Detecting that a freshly created account has cloned another account's identity markers and is now messaging that account's social graph is closer to a solved problem in applied fraud detection — one that smaller, less resourced organizations manage to perform reasonably well in narrower domains, such as financial services, every day.

This is the perceptual contradiction the chapter is named for: not that platforms fail to prevent synthetic sociality, which would be unremarkable given the genuine difficulty of operating at their scale, but that the specific pattern of what they

fail to prevent sits so awkwardly against what they otherwise demonstrably can do. The contradiction is not resolved by observing that large systems are imperfect. Large systems are indeed imperfect, in every domain, without exception. What requires explanation is not imperfection in general but the *particular shape* of this imperfection — its concentration in exactly the areas that protect ordinary users from harm, and its comparative absence in the areas that protect advertiser-facing metrics from degradation.

1.3. The Easy Answer, and Why It Falls Short

The explanation most readily available, and the one platforms themselves most often supply when pressed, runs roughly as follows: any sufficiently large, sufficiently open communicative system will attract bad actors, because openness and scale are, definitionally, the conditions bad actors require. A platform that allows anyone to create an account, post content, and reach an audience has, by the same design choices that make it valuable to legitimate users, made itself available to illegitimate ones. Fraud, impersonation, and coordinated inauthentic behavior are not failures of the system so much as inevitable features of any system built on the premises the platform was built on. Malicious actors are, in this framing, a weather condition — unwelcome, worth mitigating, but not evidence of anything about the system's priorities.

This explanation is not wrong. It is, in fact, straightforwardly true as far as it goes: an architecture that makes account creation frictionless for ordinary users makes it equally frictionless for fraudulent ones, and there is no design of a recommendation pipeline, however sophisticated, that inherently distinguishes a genuine human intention from a well-executed automated approximation of one. Both produce the same class of signal — a click, a completed view, a share — and a system built to respond to that signal class cannot, from the signal alone, tell the two apart.

But notice what the explanation quietly accomplishes beyond its literal content. It positions the platform as a neutral party to whom synthetic sociality has *happened* — a victim, in effect, of its own openness, doing its best against adversaries who exploit infrastructure built for better purposes. This framing determines, before any specific policy is debated, which kinds of response feel appropriate and which feel like overreach. If the problem is external adversaries exploiting legitimate infrastructure, the solution is better detection: more classifiers, faster takedowns, larger trust-and-safety budgets — a technical arms race requiring re-

sources but not, crucially, requiring anyone to ask why the infrastructure was built without stronger authenticity guarantees in the first place, or why it has remained that way as the scale and sophistication of synthetic activity has become increasingly visible to anyone paying attention.

That prior question — not "how do we catch more bad actors" but "why was the system built, and why does it remain, in a condition where catching them is this difficult" — is the one the easy answer is structured to avoid asking. It is also the question the remainder of this book takes as its actual subject.

1.4. What the Contradiction Is Not

It is worth being precise about what Section 1.2's puzzle does not, on its own, establish, because the temptation to overreach in the other direction is just as available as the temptation to accept the easy answer uncritically.

The contradiction does not establish that platform operators want synthetic sociality to exist, in the sense of actively preferring a fraud-saturated environment to a clean one. Nothing in the argument so far requires attributing malicious intent to any individual or team. It does not establish that every instance of platform harm traces to a deliberate choice, made by an identifiable decision-maker, to deprioritize user protection in favor of some other goal. Large organizations, as Chapter 5 will argue at length, rarely behave with the coherence that a claim of deliberate intent would require.

What the contradiction establishes is narrower and, for that reason, harder to dismiss: that the distribution of technical capability across problems is not a natural fact about which problems are hard, but a reflection of institutional priority, and that the specific priorities revealed by this distribution are worth taking seriously as data. An organization does not fail to solve a comparatively tractable problem because the problem is unsolvable. It fails to solve it because solving it was not, relative to its other objectives, worth the resources solving it would have required. That is not an accusation. It is closer to an invitation to ask what those other objectives actually are — an invitation the next two chapters take up directly, before Chapter 4 supplies the vocabulary to state the answer with precision.

1.5. The Shape of the Argument Ahead

This chapter has deliberately stopped short of explanation. Its task was to establish that the puzzle is real, that the easy answer does not fully resolve it, and that

something in the way these organizations are structured or measured is producing a specific, patterned, non-random distribution of protective failure — concentrated where protection would cost something and comparatively absent where protection is cheap or where its absence is itself profitable.

Chapter 2 examines the rhetorical move by which this distribution gets reframed as an unavoidable cost of openness rather than a consequence of design choice. Chapter 3 gives the pattern a formal name and shows precisely where, in a platform's optimization target, authenticity fails to appear. Chapter 4 generalizes the resulting structure beyond platforms entirely, and supplies the vocabulary — self-reference, reality-anchoring, the coupling coefficient ρ — that the rest of this book uses to describe the same pattern wherever it recurs: in institutions that have stopped listening to the people they harm, in minds that have stopped resolving the uncertainties they compulsively return to, and in the corrective mechanisms, examined last, that are supposed to fix all of the above and are not exempt from the same failure themselves.

The person from Section 1.1, closing the app after an ordinary evening, experiences none of this vocabulary. They experience only the accumulating weight of a low-grade suspicion they have learned, mostly without deciding to, to carry into every interaction the platform mediates. That experience is the beginning of the argument, not a side effect of it. Everything that follows is an attempt to explain, with as much precision as the subject allows, why that suspicion is not a private failure of trust but an accurate perception of a structural condition — and what, if anything, could be built so that it would no longer need to be carried.

Chapter 2

The Infrastructure Alibi

2.1. A Reasonable-Sounding Defense

Ask a platform's representatives, in almost any public forum, to account for the persistence of impersonation networks, coordinated inauthentic activity, or manufactured engagement, and a version of the same explanation reliably follows. Large-scale automation infrastructure is dual-use by nature. The systems that make legitimate recommendation, targeted advertising, and automated customer service possible are, by the same technical properties that make them useful, available to anyone who wants to abuse them. A pipeline built to distribute short video efficiently to interested audiences cannot distinguish, at the level of its own operation, between a creator sharing something they made and a coordinated network flooding the same distribution channel with recycled, algorithmically optimized filler. A system built to make account creation fast and frictionless for new users has, by that same design, made it fast and frictionless for someone creating fifty accounts to impersonate fifty different people.

This explanation is not offered cynically, in most cases, and it is not simply false. It describes something true about the architecture of large platforms: there is no layer of a modern recommendation pipeline that inherently privileges a genuine human intention over an automated, well-executed approximation of one. Both produce the same class of measurable signal. The infrastructure, examined purely as infrastructure, really does treat them alike.

2.2. What the Explanation Accomplishes

The trouble is not that this account is inaccurate. The trouble is what it is doing rhetorically, beneath its accuracy, and what it prevents a listener from asking next.

By locating the source of synthetic activity in the general-purpose, dual-use char-

acter of the underlying technology, the explanation quietly repositions the platform itself. The platform becomes a neutral substrate — value-agnostic infrastructure, like a road network or an electrical grid — upon which both legitimate and illegitimate uses happen to occur, with the platform bearing no more responsibility for the illegitimate uses than a road authority bears for the getaway vehicles that happen to use its highways. The bad actors are external. They arrived from outside, bringing their own intentions to infrastructure that was built for something else, and the platform's role is recast as the ongoing, sympathetic work of defending its own tools against people who turned them to purposes their designers never intended.

This framing does real work, and its work is not primarily descriptive. It determines, in advance of any specific policy conversation, which interventions will register as reasonable and which will register as overreach. If synthetic activity is fundamentally an external-adversary problem, then the appropriate response is a security response: better classifiers, faster takedown pipelines, larger trust-and-safety teams, closer cooperation with law enforcement. All of these are framed as matters of resourcing and technical sophistication — problems an organization can, in principle, simply spend its way toward solving, without ever having to revisit the design decisions that made the problem possible in the first place. What the framing does not invite, and structurally resists, is the earlier and more uncomfortable question: why was this infrastructure built without meaningful authenticity guarantees to begin with, and why, as the scale and visibility of synthetic activity within it has grown difficult to miss, has that fundamental design choice remained unrevisited?

2.3. The Alibi Compared to Its Object

It is useful to be precise about what kind of explanation this is, because the word *alibi* is doing specific work in the chapter's title and is not merely a rhetorical flourish.

An alibi, in its ordinary sense, does not deny that a harmful event occurred. It concedes the event and redirects attention to the question of who was where, and therefore who bears responsibility, when it happened. The infrastructure alibi operates the same way. It does not deny that impersonation networks, fraudulent marketplaces, or coordinated bot activity exist on the platform, often at considerable scale — these facts are, at this point, well documented and rarely disputed even by the platforms themselves. What it contests is the question of responsi-

bility, by locating the operative cause in the general properties of large-scale automation rather than in any choice the platform made about how that automation would be built, measured, or governed.

The comparison to an alibi also clarifies what would count as an adequate response to it, and what would not. An alibi is not defeated by re-describing the crime in more vivid detail — pointing again, more forcefully, at the impersonation account or the vanished pet deposit does nothing to the alibi's logical structure, because the alibi never denied these things happened. An alibi is defeated by establishing that the party claiming distance from the event was, in fact, positioned to prevent it and did not, or benefited from conditions that made it more likely. That is the specific argument the remainder of this book is structured to make, and this chapter is only its opening move: showing that the infrastructure explanation, however accurate about the technology, forecloses rather than answers the question of institutional choice.

2.4. A Test Case: What Gets Built and What Doesn't

The clearest way to see the alibi's limits is to compare what a platform *has* successfully built against what it claims, in the same breath, to find intractable.

The same organizations that describe synthetic-identity detection as a genuinely hard, ongoing technical challenge have, without apparent difficulty, built advertising systems capable of targeting audiences by combinations of demographic, behavioral, and inferred psychographic signal fine enough to specify a campaign down to a few hundred people in a specific geography with specific recent purchase behavior. They have built content-recommendation systems that adapt, within single-digit numbers of interactions, to a new user's previously unobserved preferences. They have built real-time bidding infrastructure that clears advertising auctions across billions of impressions in fractions of a second, accounting simultaneously for advertiser budgets, predicted engagement, and dozens of other weighted variables.

None of this is offered to suggest that identity verification at platform scale is a trivial problem — it is not, and any serious treatment of the subject should resist pretending otherwise. It is offered to make a narrower point: the claim that synthetic-identity detection remains unsolved because the underlying problem is simply too hard sits uneasily beside the observable fact that comparably hard, and in some cases harder, problems get solved routinely, whenever solving them serves an objective the organization has already decided to prioritize. Difficulty,

on its own, does not explain the pattern. What explains the pattern is a difference in what each capability is worth to the systems that would have to be built to supply it — a difference Chapter 3 gives a precise, formal shape.

2.5. Why This Matters Before the Formalism Arrives

It might seem that Chapter 2's argument is a preliminary throat-clearing on the way to Chapter 3's more rigorous claim, and in a narrow sense that is true — this chapter supplies no equations, and its argument could in principle be skipped by a reader eager to reach the formal statement of metric indifference. But the alibi is not merely a prelude to the mechanism; it is the reason the mechanism persists unexamined for as long as it does.

A formal account of why engagement metrics are indifferent to authenticity, however precise, does not by itself explain why an organization capable of understanding its own optimization target with total clarity — as these organizations plainly are, given the sophistication documented in Section 2.4 — does not simply notice the indifference and correct it. Part of the answer is structural, and Chapters 5 and 6 will take up that part directly. But part of the answer is rhetorical: an explanation is available, ready at hand, that redescribes the resulting condition as an unfortunate and unavoidable cost of scale rather than a consequence of a choice that remains open to revision. As long as that redescription is accepted — by regulators, by journalists, by users themselves — the deeper question the alibi is built to deflect does not get asked with the persistence it would otherwise require. This chapter's task has been to make sure it gets asked here.

Chapter 3

Metric Indifference

3.1. From Rhetoric to Mechanism

Chapter 2 examined a rhetorical maneuver: the redescription of synthetic sociality as an unavoidable byproduct of dual-use infrastructure, a framing that forecloses the question of institutional choice without technically denying any of the underlying facts. This chapter sets rhetoric aside and asks a narrower, more tractable question. What, precisely, do large platforms optimize, and for whom — and does the answer, stated plainly, explain the pattern Chapter 1 described without requiring any appeal to malice, negligence, or conspiracy?

3.2. The Dominant Objective

Across nearly every large consumer platform, the dominant optimization target is some variant of engagement: click-through rate, watch time, shares, comments, daily and monthly active users, session length, return frequency. These indicators did not arrive arbitrarily. Each began as a proxy for something that seemed, reasonably, to matter: if people were spending time with content, sharing it, returning to a platform repeatedly, this looked like evidence that the platform was succeeding at its stated purpose of connecting people with things they valued.

The difficulty is not with engagement metrics as originally conceived. It is with what happens once a proxy becomes a target rather than a description. Once engagement metrics were built into the objective functions of recommendation algorithms, into the quarterly reports by which public companies are evaluated by markets, and into the performance reviews by which individual engineers and product managers are evaluated by their employers, the metrics stopped functioning as an approximate description of something else and began functioning as the thing itself. This is an instance of a much older and more general observa-

tion, sometimes attributed to Goodhart, and formalized in institutional contexts by Strathern: when a measure becomes a target, it ceases to be a reliable measure of whatever it was originally introduced to represent. Once a system's behavior is organized around producing a metric, the metric's relationship to the underlying quality it once tracked can drift arbitrarily far, constrained by nothing internal to the system itself.

3.3. A Formal Statement

The dynamic can be stated with some precision. Let $E(A)$ denote the engagement produced by an activity A within a platform's environment. The optimization problem the recommendation system actually solves, in practice, is

$$\max_{A} E(A) \quad \text{subject to operational constraints.} \tag{3.1}$$

The critical feature of Equation 3.1 is what does not appear in it. If authenticity — the property of A representing genuine human expression or communication rather than automated or fabricated activity — does not enter E, then authenticity is, in the most literal sense, invisible to the objective function being maximized. An automated account that watches a video long enough to register a completion event, which then triggers a recommendation to a second account, which then generates a share action, produces exactly the same measurable signal as a human being performing the identical sequence of interactions. From the standpoint of the system solving Equation 3.1, the two are not merely difficult to distinguish. They are, for purposes of the optimization, equivalent.

This equivalence is not a description of a bug. It is a description of what Equation 3.1, correctly and successfully solved, produces. A system that maximizes $E(A)$ without A's authenticity appearing anywhere in E is not failing at its assigned task when it treats synthetic and genuine engagement as interchangeable. It is succeeding at exactly the task it was given.

3.4. The Economics of the Blind Spot

This mathematical equivalence has direct economic consequences, and tracing them clarifies why the blind spot in Equation 3.1 persists rather than being corrected as soon as it is noticed.

Advertising markets, which supply the revenue that makes free consumer plat-

forms economically viable, are concerned primarily with impressions and interaction signals — the same quantities that populate $E(A)$. An automated impression that produces a measurable interaction is, from the standpoint of the advertising transaction that monetizes it, indistinguishable from a human impression, provided the volume of automated activity does not reach a scale visible enough for advertisers to detect and penalize through reduced spending. Below that threshold, automated activity is not merely tolerated; it is, in a narrow but real sense, useful. It fills engagement gaps during periods when genuine human participation dips. It provides a steady, predictable throughput for recommendation pipelines whose performance is itself measured by the volume and consistency of the signal they generate. A system optimizing Equation 3.1 has no internal reason to distinguish a human-generated dip in engagement, which it should perhaps tolerate, from a synthetic-generated smoothing of that dip, which fills the same numerical gap.

The implication is uncomfortable, but it follows directly from the formalism rather than from any speculation about intent. Platforms do not merely fail to eliminate synthetic activity because eliminating it is technically difficult, although it may be. They operate within a set of economic conditions in which synthetic activity, up to some threshold, produces signals their own business model treats as unambiguously positive. The relationship between a platform and the synthetic activity occurring within it is therefore not straightforwardly adversarial, in the way the infrastructure alibi from Chapter 2 implies. It is, at minimum, a relationship of mutual indifference — and in specific, measurable respects, one of mutual benefit, at least up to the point where visibility to advertisers or regulators changes the calculation.

3.5. Indifference Without Malice

It is worth restating, with more force than a single caveat usually receives, that none of this requires deliberate deception on the part of any individual or team. The mechanism described in this chapter does not depend on a executive, in some private meeting, deciding that fraud is an acceptable cost of doing business. It requires only the much more mundane sequence already described: a system is built to maximize a measurable signal; a class of activity is discovered, or arises spontaneously, that produces that signal without the underlying property the signal was originally meant to indicate; and no correction mechanism intervenes, because no correction mechanism was built into the objective function in the first

place, and because — per Chapter 2 — the resulting condition is available to be redescribed as an unavoidable feature of scale rather than a symptom worth investigating.

The absence of malice does not diminish the structural condition Equation 3.1 describes. A system indifferent to authenticity will, regardless of what any individual within the organization believes or intends, produce an environment in which authenticity goes unprotected. This is the chapter's central claim, and it is worth holding in mind exactly as stated, without either softening it into a claim about corporate villainy it does not make, or dismissing it as a claim about mere technical limitation, which understates what Equation 3.1 actually shows: not that platforms cannot detect synthetic activity, but that nothing in what they are built to maximize requires them to.

The word "indifferent," however, is doing more work here than the argument so far has earned, and it is worth flagging the simplification before moving on rather than letting it stand as the chapter's final word. Equation 3.1 shows that authenticity does not appear in the objective a recommendation system is built to maximize; it does not, on its own, establish that the organization operating that system is indifferent to authenticity in any broader sense. Large platforms do fund trust-and-safety teams, employ content moderators by the thousands, and report real expenditure on exactly the harms this book is concerned with — facts a purely indifferent organization would have no reason to produce. The gap between "authenticity is absent from this equation" and "the organization is indifferent to authenticity" is real, and this chapter has, so far, moved across it faster than it should. Chapter 5 takes up this gap directly, replacing the single-agent "indifference" of this chapter with a more accurate account of what happens when locally rational subsystems — a recommendation team optimizing Equation 3.1, a trust-and-safety team optimizing something else entirely, with neither fully accountable to the other — produce, in aggregate, a result that looks like indifference from the outside without any single part of the organization actually holding that attitude. "Indifference" is the right word for what Equation 3.1 shows about a recommendation system in isolation. It is not yet the right word for the organization that built it, and readers should hold the claim provisionally until Chapter 5 supplies the more complete picture.

3.6. What Remains Unexplained

Three chapters into this book, a skeptical reader is entitled to ask what, precisely, has been established beyond a redescription of an already-familiar complaint about engagement-driven design. The answer is that Chapters 1 through 3 have, so far, each supplied a piece of a single pattern without yet naming the pattern itself. Chapter 1 established that the pattern is real and non-random. Chapter 2 established that a ready explanation exists whose function is to prevent the pattern from being investigated as a matter of institutional choice. This chapter has established, with as much formal precision as the subject currently allows, exactly where in a platform's objective function the pattern originates, and why solving Equation 3.1 correctly is sufficient to produce it, with no further assumption required.

What remains — and what the next chapter undertakes — is the harder task of showing that this is not a fact about platforms specifically, or about engagement metrics specifically, but an instance of something that recurs wherever an evaluating system's objective fails to include a channel it cannot control. Equation 3.1 is not the general case. It is the first, and clearest, example of it.

Chapter 4

The Geometry of Correction

4.1. Three Instances in Search of a Pattern

Chapters 1 through 3 described a single organization from three angles. The platform experiences a perceptual contradiction: it commands some of the most sophisticated predictive infrastructure ever built, and yet the ordinary user's experience of it is one of impersonation, fraud, and manufactured consensus. The platform offers an alibi: it describes the infrastructure enabling this condition as neutral, general-purpose, and therefore not properly its responsibility. And the platform exhibits a specific mathematical property: its optimization target, engagement, does not contain authenticity as a variable, so authenticity's absence or presence makes no difference to what the system does.

These are not three separate observations that happen to concern the same company. They are the same observation, examined at three different resolutions. The purpose of this chapter is to state that observation once, in a form general enough that the reader will recognize it again — without prompting — in Part II, where the unit of analysis becomes the institution rather than the platform; in Part III, where the unit of analysis becomes the individual mind; and in Part IV, where the question becomes what can be built to prevent it.

4.2. Self-Reference and Reality-Anchoring, Defined

Let a system's behavior be governed by an evaluation function

$$E : X \to \mathbb{R}$$

over a space of possible states X. Every functioning system of the kind this book is concerned with — a recommendation algorithm, a bureaucracy, a market, a

mind — produces, as a byproduct of operating, a stream of internal signals about its own activity. Call this internal signal set $S(x)$: clicks and watch time for a feed, citations for a research community, symptoms and fantasies for a psychic apparatus, quarterly reports for a firm.

A system is **self-referential** when its evaluation depends entirely on this internal signal set:

$$E(x) = f\big(S(x)\big).$$

A system is **reality-anchored** when its evaluation also depends on some channel $O(x)$ that the system does not fully author or control:

$$E(x) = f\big(S(x), O(x)\big).$$

The word "channel" is doing specific work here and is worth pausing on. O is not "reality" in some diffuse sense. It is a *measurable point of contact* between the system's evaluation and something the system cannot simply generate more of by trying harder at what it already does. A pull request either compiles against the existing codebase or it does not; the codebase is not a signal GitHub invents, it is a constraint GitHub's evaluation is forced to consult. A prediction either matches a subsequent observation or it does not; the observation is not conjured by the forecaster. A claim of romantic reciprocation either survives contact with the other person's actual behavior or it does not; the other person is not a construct of the one who desires them, however much the predictive apparatus described in Chapter 10 might wish otherwise.

We can express the degree to which a system remains answerable to O as a coupling coefficient:

$$\rho = \frac{\partial E}{\partial O}.$$

This is deliberately framed as a coefficient rather than a switch. Few real systems are purely one or the other. What varies — across systems, and within a single system over time — is how strongly a change in the external channel is permitted to move the evaluation. When $\rho \to 0$, O has stopped mattering to E in practice, whatever the system's design documents or mission statements claim about its intentions. When ρ remains bounded away from zero, the system's self-assessment

can still be overruled by something it did not produce.

Chapters 1 through 3, restated in this vocabulary: Chapter 3's Equation 1 is the formal claim that a platform's engagement objective has $\rho \approx 0$ with respect to authenticity — the partial derivative of engagement with respect to whether an interaction was genuine is, for practical purposes, zero. Chapter 2's alibi is the rhetorical strategy by which an organization explains away a low ρ as an unavoidable feature of scale rather than a consequence of what it chose to measure. Chapter 1's perceptual contradiction is what it feels like, from the inside, to be a participant in a system whose ρ has fallen without anyone deciding that it should.

4.3. A Composite Channel

It would be a mistake to imagine O as always a single, unified signal. Different systems anchor to different — and sometimes multiple — external channels, and the mechanisms that raise ρ in one domain will not necessarily transfer to another. A scientific claim is anchored by replication and by predictive success, which are related but not identical channels; a claim can survive one and fail the other. An identity system, as later chapters will argue, can be anchored by a persistent, non-erasable event history — one kind of channel — and separately by the requirement that its interpretations remain within reach of a specific human participant's perspective — a different kind of channel entirely, doing different work.

Nothing in the definition of ρ requires O to be scalar. Treat it, where useful, as composite: $O = (O_1, O_2, \ldots)$, with the system's overall reality-anchoring depending on its coupling to each. This matters because a design intervention that raises ρ with respect to one component of O may leave the system fully self-referential with respect to another. A platform that verifies identity provenance has raised ρ with respect to one channel — who is speaking — while leaving ρ with respect to a different channel — whether what is said is true — entirely untouched. The two problems are often conflated and require different remedies.

4.4. The Correction Suppression Principle

A natural but mistaken intuition is that low ρ means reality has gone missing — that self-referential systems are self-referential because the external world has somehow stopped generating signals. This is very rarely what happens, and the mistake is worth naming precisely because so much of what follows in this book depends on not making it.

Consider the distinction between two conditions:

$$O(x) = \varnothing \qquad \text{versus} \qquad O(x) \neq \varnothing \text{ but } \frac{\partial E}{\partial O} \approx 0.$$

In the first condition, no external channel exists at all — there is genuinely nothing to consult. This is rare. In the second condition, the channel exists, continues to produce information, and the system has simply stopped letting that information move its evaluation. The bridge does not stop being a physical structure subject to load-bearing constraints because an engineering firm's incentive structure has drifted; it either holds or it collapses regardless of what anyone measured. The software does not stop crashing because a product team's dashboard rewards ship velocity over stability. The rental deposit is still stolen, the impersonated friend's contacts are still deceived, the recommendation system's user is still worse off by their own subsequent judgment — whether or not any of these facts register anywhere in the evaluating system's objective function.

This gives us a principle worth stating on its own terms, since it recurs often enough in the chapters ahead to warrant a name:

> ***The Correction Suppression Principle.*** *Pathological self-reference typically arises not because corrective channels cease to exist, but because evaluative systems become progressively insensitive to channels that remain, in fact, fully present.*

The practical consequence of this principle is that the diagnostic question is almost never "does an external check exist?" It is "is the system's evaluation still required to pass through that check?" These call for different remedies. Where O is genuinely absent — a domain in which no one has yet built a way to measure the thing that matters — the task is measurement design. Where O is present but unconsulted, measurement design accomplishes nothing; the task is forcing consultation, which is a different and often harder problem, because it usually means changing who benefits from *not* consulting a channel that everyone agrees exists.

Later chapters will return to this distinction directly. The externalized costs of Chapter 7 are not costs that occur in the absence of a corrective channel — the harmed party is, in nearly every case, screaming directly into a channel that exists and is simply weighted at zero. The jouissance of Chapter 11 is not a condition of the real having vanished from a subject's experience; it is a condition of the real persisting, unresolved, precisely *because* the predictive apparatus keeps returning

to it without letting it close. The productive friction of Chapter 13 does not manufacture an external channel that wasn't there before; it restores the requirement that the channel already present be consulted before the system proceeds. In every case, the remedy is not invention. It is re-coupling.

4.5. Recursion: Who Corrects the Corrector?

The framework as stated so far treats E, S, and O as properties of a single system under evaluation. But nothing prevents the same apparatus from being applied to a system whose entire purpose is to serve as somebody else's O.

Suppose a system S_1 is evaluated by E_1, and — recognizing that S_1's ρ has fallen, by whatever diagnostic Part II or III supplies — we introduce a corrective system C, whose function is to evaluate S_1 and hold it answerable to some channel it had stopped consulting. This is, in the most general terms, what a regulator does, what an auditor does, what governance in general purports to do, and it is the implicit hope behind most proposals — including some that will appear later in this book — for fixing systems that have drifted toward self-reference.

The observation this section exists to make is simple, and it is easy to miss precisely because C is introduced with corrective intent: C is itself a system. It possesses, whether its designers thought about it explicitly or not, its own evaluation function E_C, its own internally generated signals S_C — audit completion rates, enforcement statistics, cases closed — and its own relationship, strong or weak, to some external channel O_C that would tell it whether its corrections are actually working. Because C is a system of the same general kind as S_1, it is describable by the same coefficient:

$$\rho(C) = \frac{\partial E_C}{\partial O_C}.$$

There is no principled reason $\rho(C)$ should be assumed high simply because C was built with corrective purpose. A regulator can come to evaluate itself by enforcement volume rather than by whether the underlying harm decreased. An internal trust-and-safety function, introduced in Chapter 6, can come to evaluate itself by policy violations processed rather than by fraud actually prevented. An auditor can come to evaluate itself by audits completed rather than by whether the audited system changed. Correction, in other words, is not a state a system arrives at and then occupies permanently. It is itself subject to the dynamic this chapter has been describing, and a corrective system with $\rho(C) \to 0$ is not a correction at

all — it is a second self-referential system wearing the first one's uniform.

This is not a caveat appended to the framework developed above. It is the framework's most consequential result, and it changes the kind of question the rest of this book is entitled to ask. Without recursion, ρ is a diagnostic — a way of measuring whether a given platform, institution, or mind has drifted toward self-reference. With recursion, ρ becomes something closer to a constitutional principle, because it applies without exception to whatever is proposed as the remedy. The question is no longer only *is this system self-referential?* It is also *is the mechanism built to correct it self-referential?* — and then, by the same argument, applied again to that mechanism's own overseer, for as many layers as anyone cares to introduce. Regulators, auditors, peer review, alignment boards, governance committees, and every institution proposed in Part IV of this book are all, on this analysis, systems of the same kind as the platform in Chapter 1, and none of them is exempt from being asked the question it was built to ask of something else.

State the result plainly, because it is the sentence this chapter is built to deliver: **no corrective system stands outside the geometry of correction.** A correction that has forgotten this fact about itself is not a correction in progress. It is the pathology, recurring one level up. Part IV does not attempt to escape this consequence by proposing a final, exempt authority — no such proposal would be credible after this chapter — but by asking what kinds of corrective structures remain viable once no permanent exemption is available to any of them. The regress this implies — a corrector requiring its own corrector, indefinitely — is not a loose end this book failed to tie off. It is the consequence the framework predicts, stated once here so that its recurrence at the end of Chapter 17 reads as confirmation rather than as an admission of defeat.

4.6. The Boundary of the Claim

A framework stated this generally invites two opposite misreadings, and it is worth fixing the boundary of what has actually been established before the instantiations begin.

The claim is not that all self-referential systems are illegitimate, or that automation itself is the pathology. It is a claim about ρ: the coupling of an evaluating system to a channel outside its control. Chapter 9 will show that a user's informed, consenting engagement with a synthetic actor sits entirely outside this claim's scope — the geometry of correction concerns the evaluating system's own coupling, not the composition of $S(x)$, and synthetic content inside $S(x)$ is not by

itself evidence of anything. The framework has nothing to say, one way or the other, about automation that both parties know to be automation.

Nor is this a claim about any particular theory of mind. O is defined here only as an external channel a system does not fully author — general enough to be instantiated many ways. Chapter 10 will instantiate it specifically as "the real," in a technical and psychoanalytic sense, and will draw on the Free Energy Principle to describe the mechanism by which precision-weighting can be captured by an external system. That is one admissible instantiation among others, introduced where it does its work and nowhere before. A reader who does not accept that particular cognitive-scientific apparatus loses nothing in the chapters that do not depend on it.

And this is not a claim that Spherepop, quadrangulation, or any other mechanism proposed later in this book is *the* solution. Section 4.5 has already ruled out the possibility of an exempt, final corrective authority; what remains is a design space of candidate mechanisms, each judged by the same standard this chapter applies to everything else — its own coupling to something it does not control.

One final clarification belongs here, concerning the evaluative language the rest of this book uses freely — words like "hollow," "blind spot," and "concentrated power," none of which are neutral. The distinction between self-reference and reality-anchoring, exactly as defined above, is descriptive: ρ measures a structural property of an evaluating system, and nothing in that measurement, by itself, says whether a given position on the axis is good or bad. A hermit's private evaluation of their own garden can be as self-referential as it likes without this book having any complaint to register. The book's evident and repeated preference for high ρ is a separate, substantive commitment, not a hidden implication of the formalism, and it is worth stating once, plainly, rather than letting it operate silently underneath descriptive-sounding words for the next thirteen chapters. The commitment is this: reality-anchoring matters, in the cases this book actually examines, because the systems in question affect people who did not design them, did not consent to their objective functions, and cannot simply exit — platform users subject to a recommendation algorithm they had no part in specifying, employees governed by institutions whose incentive structures were set above them, and, in Part III, subjects whose own predictive architecture is recruited by an environment they did not build and in most cases cannot fully see. Low ρ becomes this book's concern specifically where the consequences of a system's drift fall on people standing outside the boundary of the system doing the evaluating. That is

the normative claim underneath the descriptive apparatus, and the book is more honest for saying so directly than for letting it leak through unexamined in the choice of adjectives.

One further limit is worth stating plainly, since it is the kind of gap a careful reader will notice whether or not this book admits it. Nothing in this chapter specifies how ρ is to be estimated for a real system from available evidence — what data an auditor would gather, what procedure would convert that data into a number rather than a direction, or how confidently two analysts examining the same institution should expect to agree on where it falls. ρ is introduced here as a diagnostic orientation: a way of asking whether a system's evaluation is drifting toward or away from some external channel, and of recognizing the direction of that drift from the pattern of a system's behavior over time — which channels it demonstrably responds to, and which it merely claims to. It is not, in this book, an operationalized measurement with a specified estimator attached, and the difference between those two things is not small. A reader equipped only with this chapter could not walk into an organization and return with a number for its ρ. What they could return with is a sharper question than they had before, applied consistently across every domain the rest of this book examines. That diagnostic sharpening, rather than a finished instrument, is what is being offered here; building the instrument is a separate undertaking this book does not attempt.

What this chapter does establish, and establishes fully, is the following: that self-reference and reality-anchoring are two ends of a single graduated property, applicable to any evaluative system regardless of substrate; that a system's position along that axis can drift under ordinary optimization pressure without any actor deciding to abandon correction; that the drift is diagnosable independently of domain, whether the system in question is a platform, an institution, or a mind; and that no corrective system — including any this book will go on to propose — is exempt from the same scrutiny. These four claims are what the rest of the book is entitled to assume. Everything else is instantiation.

Part II

Institutional Self-Reference

Chapter 5

Against the Unified Maximizer

5.1. A Convenient Fiction

Chapter 3 wrote a platform's behavior as the solution to a single optimization problem: maximize $E(A)$ subject to operational constraints. That formalism was useful, and nothing in this chapter retracts it. But it also flattened something the coming chapters need to recover, because the flattening, left uncorrected, produces a specific and consequential misdiagnosis.

Writing a platform's behavior as $\max_A E(A)$ implies, even if only by the economy of notation, that some coherent agent is doing the maximizing — a single "the platform" that surveys its options and selects the one that best serves a unified objective. This is a convenient fiction, useful for isolating the mechanism Chapter 3 needed to demonstrate, but it is not a description of how any large technology company actually operates, and mistaking the fiction for the reality leads directly to bad diagnoses and worse remedies.

5.2. What Large Organizations Actually Are

A large platform company is not a single mind with a single objective function. It is a bureaucracy — in the strict, structural sense of the word, not the pejorative one — composed of tens or hundreds of thousands of people distributed across divisions with different mandates, different time horizons, different performance metrics, and, not infrequently, directly competing incentives.

Consider the range of teams whose combined behavior produces what an outside observer experiences as "the platform." An engineering team responsible for recommendation infrastructure is evaluated, promoted, and funded according to measurable improvements in model performance against engagement metrics — precisely the metrics formalized in Chapter 3's Equation 3.1. A trust-and-safety

team is evaluated according to a largely separate set of indicators: policy violations detected, accounts actioned, response time to reported abuse. A growth or product team is evaluated by user acquisition and retention figures. A policy team is evaluated by its success in managing regulatory relationships and public controversy, which follows a political rather than an engineering logic entirely. A legal team is evaluated by litigation exposure. Each of these teams optimizes, locally and often quite skillfully, for the objective it has been assigned. The organization's aggregate behavior is the sum of these locally optimized subsystems interacting — but that sum has no reason to exhibit the coherence one would expect from a single agent solving a single, well-specified problem, because no such agent, and no such problem, exists anywhere inside the organization actually generating the behavior.

5.3. Why the Distinction Changes the Diagnosis

This matters because it changes what kind of explanation is available for the persistence of synthetic sociality, and therefore what kind of remedy is worth proposing.

If a platform behaved as a coherent maximizer, the persistence of fraud and impersonation despite obvious technical capability elsewhere would most plausibly indicate a deliberate strategic choice: someone, somewhere, decided that the marginal engagement supplied by tolerating a certain level of synthetic activity was worth more than the reputational and user-trust cost of eliminating it, and the organization has acted accordingly. This is not an unreasonable hypothesis, and it may in fact be true of specific decisions at specific companies at specific times. But it is not the only available explanation, and treating it as though it were the only one forecloses a different, more dispersed diagnosis that is, in many documented cases, closer to what actually happens.

The more dispersed diagnosis runs as follows. The team with the mandate to address coordinated inauthentic behavior may have meaningfully fewer resources, relative to its problem's difficulty, than the teams building the infrastructure that makes such behavior possible at scale — not because anyone decided detection should be underfunded, but because detection's contribution to the metrics that determine internal resourcing is harder to demonstrate than infrastructure's contribution to the metrics that determine internal resourcing. Detecting a sophisticated synthetic network may require capabilities and data access that cross organizational boundaries the company's internal structure was never designed to

bridge, so the problem falls, quite literally, into the gap between two teams' mandates, belonging fully to neither. The dashboards by which senior leadership tracks platform health may simply not include indicators of synthetic activity prevalence, not because such indicators were deliberately excluded, but because building and maintaining them was never anyone's assigned responsibility, and things that are no one's assigned responsibility tend not to get built, at any organization, for reasons that have nothing to do with the specific harms at stake.

5.4. Explanation, Not Excuse

It is important to be precise about what this chapter's argument does and does not accomplish, because the distinction between explanation and excuse is easy to blur here and the blurring serves no one.

Reframing platform behavior as the emergent product of distributed, locally rational optimization rather than the deliberate output of a coherent strategic actor does not exonerate the resulting harm. The impersonation account still deceives its target's contacts whether its persistence traces to a boardroom decision or to a resourcing gap between two teams that were never asked to coordinate. The rental deposit is still stolen either way. What the distributed account changes is not whether the outcome is bad, but what kind of intervention is capable of fixing it.

An organization that has made a coherent strategic choice to tolerate fraud because fraud inflates metrics can, in principle, be moved by pressures that change the calculus of that choice: regulatory penalties large enough to outweigh the metric gain, advertiser defection triggered by reputational damage, sustained public pressure that raises the political cost of inaction. These are the levers most public discussion of platform accountability instinctively reaches for, and they are not useless — but they are calibrated to a coherent-agent diagnosis, and their effectiveness depends on that diagnosis being correct.

An organization whose failure to protect users traces instead to distributed organizational dysfunction requires a different kind of intervention, one aimed not at changing a strategic calculation but at closing structural gaps: internal measurement systems that make synthetic activity visible to the decision-makers who currently cannot see it; resourcing and promotion criteria that reward cross-boundary problem-solving rather than only locally measurable performance; organizational structures that assign clear ownership to problems that currently fall between mandates. Applying the first kind of remedy — public pressure aimed at a strate-

gic decision-maker — to the second kind of problem accomplishes little beyond generating a public statement of commitment that has no internal mechanism obligated to enforce it. This is, not coincidentally, a familiar pattern: many platform commitments to "doing better" on trust and safety amount to exactly this mismatch, a coherent-agent remedy applied to a distributed-system problem, and their frequent failure to produce measurable change is precisely what the mismatch predicts.

5.5. A Note on the Coupling Coefficient

It is worth pausing to connect this chapter explicitly to the vocabulary introduced in Chapter 4, since the connection is not merely decorative.

Chapter 4 defined ρ as a property of a system's evaluation function — the degree to which that evaluation remains coupled to some external channel O it does not control. Nothing in that definition required the evaluating system to be a single coherent agent. A distributed organization, composed of many locally optimized subsystems, can be described as possessing an aggregate ρ with respect to any given channel, even though no individual subsystem within it is solving a unified optimization problem. Indeed, the distributed case may often produce a *lower* aggregate ρ than a coherent-agent case would, for a specific structural reason worth naming: in a coherent agent, a single decision-maker who notices that some important channel O is being ignored can, in principle, simply correct the objective function directly. In a distributed system, correcting ρ requires that some subsystem be assigned responsibility for the channel in question, resourced to monitor it, and empowered to act on what it finds — three separate organizational conditions, any one of which failing is sufficient to leave ρ near zero indefinitely, with no individual decision anywhere in the system responsible for that outcome.

This is the institutional-scale version of a point Chapter 6 develops further: that the absence of a single, identifiable villain does not imply the absence of a structural cause, and that some of the most durable instances of $\rho \to 0$ are durable precisely because no coherent agent exists whose decision to change course would be sufficient to fix them.

Chapter 6

Distributed Responsibility, Concentrated Power

6.1. An Asymmetry Worth Naming

Chapter 5 established that large platform companies behave as distributed optimization systems rather than coherent maximizers, and that this distribution partly explains why synthetic sociality persists without requiring any single decision-maker to have chosen it. This chapter examines a specific governance pattern that compounds that dynamic, and does so in a way that is not incidental to how these companies happen to be run, but characteristic of how a particular class of technology firm has come to be structured.

Call it the portfolio model of technical leadership: operational responsibility distributed widely across semi-autonomous divisions, each pursuing its own mandate with considerable independence, while ultimate strategic authority remains tightly concentrated at the top of the organization — frequently in a single founder, whose formal control is reinforced through voting-share structures that make that authority largely immune to the diffusion of risk the rest of the organization is subject to.

6.2. The Managerial Rationale

The portfolio model is not adopted arbitrarily, and it is worth taking its stated justification seriously before examining its consequences. Distributing development work across multiple, semi-independent teams allows an organization to pursue several approaches to a problem in parallel, reduces its dependence on the judgment of any single technical leader, and avoids the bottlenecks that centralized, hierarchical decision-making tends to produce as an organization scales. In domains where the technical landscape is genuinely unpredictable — frontier machine learning research is the clearest current example — parallel experimen-

tation with a high tolerance for individual failure is a defensible strategy for managing uncertainty that a more centralized structure would handle poorly.

None of this is in dispute. The managerial case for distributing operational responsibility is sound on its own terms, and this chapter is not an argument against it.

6.3. The Governance Consequence

The argument concerns what happens to accountability once operational responsibility has been distributed this way while strategic authority has not been distributed at all.

When responsibility is spread across many semi-autonomous groups, but the ultimate authority to set direction, allocate resources, and determine what the organization is for remains concentrated in a small number of hands — often, in founder-led technology firms, effectively one — a characteristic institutional geometry emerges. When an initiative succeeds, the success is readily attributed upward: to the vision, judgment, or foresight of concentrated leadership, which set the direction that made the success possible. When an initiative fails, or when a systemic harm emerges that no single team's mandate covers, responsibility is distributed laterally instead, across a network of teams whose individual accountability relationships are genuinely, and often deliberately, difficult to trace. Success climbs; failure spreads. This is not a conspiracy in the sense of deliberate coordination toward that outcome — it is simply what the combination of concentrated authority and distributed operational responsibility produces, mechanically, without anyone needing to intend it.

6.4. Why Synthetic Sociality Fits This Pattern Exactly

The relevance to this book's central concern is direct. Consider the specific harms Chapter 1 opened with: the impersonation network that exploits an easy account-creation pipeline; the coordinated bot farm that games an engagement algorithm; the fraudulent marketplace listing that exploits the absence of provenance verification. Each of these is, individually, addressable by a specific team with a specific mandate — the account-integrity team, the anti-abuse team, the marketplace-trust team. But the *systemic condition* that makes all three simultaneously viable and durable — the underlying design choices around identity verification, provenance tracking, and cross-team coordination discussed in Chapters 2 and 5 — belongs to no team's mandate. It is not that no one is responsible for it. It is that

responsibility for it was never assigned to anyone, because assigning it would require an act of concentrated strategic authority that the same governance structure protecting that authority from operational risk also, structurally, tends to protect from having to exercise itself in this way.

This is why the argument against treating the platform as a unified maximizer, developed in Chapter 5, does not resolve into a simple absolution of leadership. It does the opposite: it complicates the available remedies considerably. Changing the objective function of a genuinely coherent organization — persuading or compelling a single strategic decision-maker to weigh authenticity more heavily — is difficult but conceptually straightforward; it is the kind of change public pressure, regulation, or market discipline is reasonably well suited to producing. Restructuring the accountability relationships of an organization whose behavior emerges from many semi-independent teams, each locally rational, whose ultimate authority is structurally distanced from the operational consequences those teams produce, requires intervention at a different level entirely — one that does not stop at internal reform, because internal reform in this structure has no natural mechanism to sustain itself against the same governance asymmetry that produced the problem, but instead requires regulatory frameworks capable of imposing accountability at the specific point where strategic authority and operational impact actually converge.

6.5. A Second Instance of the Correction Suppression Principle

This chapter's argument is a direct institutional instance of the Correction Suppression Principle named in Chapter 4, and it is worth stating the instance explicitly rather than leaving the reader to infer it.

The channel is not absent. Users experience the harm; some of them report it; journalists document it; researchers study it; in some jurisdictions, regulators investigate it. O, in the notation of Chapter 4, is not empty. What is absent is a structural mechanism by which that channel's signal reliably reaches the one part of the organization — concentrated strategic authority — actually positioned to change the design choices generating the harm, without first passing through a lateral network of divided operational responsibility that has every local incentive to treat the signal as somebody else's problem. The suppression here is not of the signal itself, which continues to arrive. It is of the *coupling* between that signal and the part of the system with the power to act on it — which is exactly the distinction Chapter 4 insisted mattered, and exactly the reason a remedy aimed

at generating more signal (more user reports, more journalism, more documentation) accomplishes comparatively little on its own, without also addressing the structural reason that signal, once generated, does not reliably travel to where a decision could be made in response to it.

6.6. Toward Chapter 7

If Chapter 5 explained why no one may have decided to tolerate synthetic sociality, and this chapter has explained why the governance structures common to large platform firms make that toleration durable even after it becomes visible, the question that remains is who actually bears the resulting cost, and how that cost is distributed once it has been, in effect, structurally orphaned within the organization that produced it. That is the subject of Chapter 7.

Chapter 7

The Externalization of Cost

7.1. From Structure to Consequence

The preceding two chapters have been, in a specific sense, structural: they explained why synthetic sociality persists without requiring a villain, by describing the organizational conditions — distributed optimization, concentrated authority insulated from operational consequence — that make the pattern durable once it emerges. This chapter turns from structure to consequence, and asks a more direct question. Who pays for this, and what does the shape of that payment reveal about the correction channel Chapter 6 showed to be present but structurally unreachable?

7.2. A Catalogue, Briefly

It is worth being concrete, because the abstraction of the preceding chapters can obscure the specificity of what is actually being described. Impersonation accounts clone a person's photograph, name, and fragments of their public history to construct a convincing duplicate, which is then used to solicit money from that person's actual contacts or to distribute fraudulent links to their actual audience — exploiting, with some precision, the trust that an established relationship implies. Fake rehoming or adoption listings collect deposits from people seeking to adopt an animal that does not exist, or is not available, with the operator disappearing and reappearing under a new listing in a different regional community group once the first is reported. Rental scams collect deposits on apartments that are not for rent, or do not exist, trading on the platform's implied guarantee of community legitimacy. Investment fraud accounts construct apparent credibility through stolen identity signals before promoting schemes designed to extract money from an audience that believed it was following someone real.

None of these are sophisticated attacks in any technical sense. They exploit elementary and, in each case, well-documented features of platform design: frictionless account creation without meaningful identity verification, the absence of durable provenance or history signals for new accounts, the capacity to operate across jurisdictions and communities without accountability to any single one of them, and algorithmic amplification that can place new, unearned content in front of large audiences before any reputation — good or bad — has had time to attach to it.

7.3. An Asymmetric Distribution

The cost of this activity is not distributed evenly, and the pattern of its distribution is itself informative. It falls, in the overwhelming majority of cases, on individuals — and specifically on individuals who are least equipped, in time, expertise, or institutional support, to absorb it. A person who has lost a rental deposit to a fabricated listing must determine, alone, whether pursuing recovery is worth the time it will cost, against uncertain odds of success, through channels — police reports across jurisdictions, platform reporting systems with inconsistent response times, small-claims processes not designed for this kind of harm — that were not built with this specific injury in mind. A person whose identity has been cloned must navigate a reporting process that may move slowly or inconsistently, while the duplicate account continues, in the interim, sending fraudulent messages to that person's actual friends and colleagues, a harm that compounds for every day the reporting process takes.

The platform's own institutional response to a confirmed instance of this harm is, characteristically, to remove the offending account. This is not nothing, but it is worth being precise about what it does and does not accomplish. It removes a single node. It does not touch the network behind that node, which typically persists, redeploys under a new account, and resumes — sometimes within hours. The cost of building a new account is, per Chapter 2's argument, close to zero; the cost of investigating, reporting, and recovering from the harm that account caused is, for the individual who bore it, considerable and largely non-recoverable.

7.4. Historical Continuity and a Genuine Novelty

It would be a mistake to conclude from the preceding sections that this asymmetry is entirely new, and worth pausing on the ways it is not, before turning to the one respect in which it genuinely is.

Anxiety about the authenticity of mediated communication has accompanied nearly every significant communication technology to reach mass adoption. The spread of print raised contemporaneous concern that the sheer proliferation of available texts would overwhelm ordinary readers' capacity for critical evaluation. Broadcast media was accused, across a substantial critical literature, of manufacturing false consensus and homogenizing culture in place of genuine political participation. Early internet forums generated their own extensive discussion of the unreliability of anonymous communication and the practical impossibility of verifying an interlocutor's identity or sincerity. In each case, the underlying worry was structurally similar to the one this book is concerned with: that a new medium would displace authentic contact with a cheaper, more easily manufactured substitute, and that human communities would find themselves populated by performances rather than persons. And in each case, the worry was neither wholly vindicated nor wholly refuted — new media did introduce genuinely new mechanisms of manipulation, and human communities did, over time, develop new norms, literacies, and institutions that allowed meaningful interaction to persist alongside the degraded forms the new medium also made possible.

This continuity matters, because it should discourage the book's argument from overclaiming novelty where none exists. But one feature of the present condition is not simply the latest iteration of a recurring anxiety, and it is worth isolating precisely, because later chapters — Chapter 16 in particular — build a specific design response around it. Earlier mass media manipulated at scale, but did so without personalization: a propaganda campaign addressed an undifferentiated public with the same message. The synthetic activity described in Section 7.2 does something categorically different. An impersonation account that clones a specific person's identity and then operates *inside that person's actual social graph* — sending messages to their actual friends, colleagues, and family — is not addressing an undifferentiated public. It is exploiting the intimacy of an established relationship rather than the diffuse attention of a mass audience, and it is doing so at a cost, per account, low enough to be repeated across thousands of individual social contexts simultaneously. That specific combination — personalized infiltration into a real, pre-existing trust network, generated at a marginal cost approaching zero — has no clean precedent in the earlier communication regimes this section has just surveyed, and it is the reason the design response developed in Part IV of this book anchors specifically to the preservation of a *particular* human participant's perspective, rather than to some more general, undifferentiated notion of a reality-check.

The Geometry of Correction

7.5. The Suppression Principle, Once More

The pattern this chapter has documented is, again, an instance of the Correction Suppression Principle from Chapter 4, and it is worth closing on the instance explicitly, since it will recur once more in Chapter 17's discussion of governance.

The harmed individual, in every case catalogued in Section 7.2, is not a silent victim of an absent channel. They are, characteristically, an unusually motivated and specific source of exactly the signal that would, if it reliably reached the part of the system positioned to act on it, justify a design change: they know precisely what happened, when, and how; they often report it through the mechanisms the platform provides for that purpose. O is present, specific, and in most cases directly volunteered. What Chapter 6 already established is that this signal, once volunteered, is structurally unlikely to travel from the individual report to the concentrated strategic authority capable of changing the underlying design — not because the report was ignored by any particular person, but because no part of the distributed organization Chapter 5 described has both the mandate and the incentive to aggregate thousands of individually reported harms into the kind of systemic case that would move a design decision at the top. The cost, in other words, is not merely externalized onto the individual in the financial or emotional sense this chapter has documented. It is externalized in the more specific sense that the individual is left to carry, entirely alone, the burden of a correction channel that the institution receiving their report has no structural mechanism to actually use.

Part III

Subjective Self-Reference

Chapter 9

Desire and Synthetic Interaction

9.1. A Necessary Firewall

Before Part III turns to the cognitive mechanisms that make synthetic sociality so effective at capturing attention, it is worth stopping to close off a misreading that the argument so far makes tempting, and that would, if left uncorrected, undermine both the diagnosis already offered and the design proposals still to come.

Nothing in Chapters 1 through 7 licenses the conclusion that synthetic or automated interaction is itself the problem. That conclusion does not follow from the argument, and this chapter exists specifically to make sure it is not drawn.

9.2. What People Actually Choose

Consider how much of ordinary digital life already consists of engagement with actors both parties understand to be automated, scripted, or otherwise non-reciprocal, entered into knowingly and without complaint. A parasocial relationship with a content creator who posts on an algorithmically optimized schedule, whose visible warmth toward an audience of millions cannot be individually reciprocated in the way a friendship can, is nonetheless experienced by many people as a genuinely valuable, freely chosen form of engagement. A brand account that responds to customer messages through an automated system is understood by nearly everyone who interacts with it to be automated, and is used anyway, because the interaction serves its purpose regardless. An entertainment stream produced by a coordinated content team behind a single on-screen persona is enjoyed by audiences who are, in the overwhelming majority of cases, entirely aware that a team rather than a solitary individual is responsible for what they are watching.

These are not marginal or reluctant uses of platform technology. They are widely

and repeatedly chosen, and they are often attractive precisely because of the properties that might, on a naive reading of this book's argument so far, seem disqualifying: the engagement is one-directional, the social obligation is minimal or absent, and the experience of connection is available without the reciprocal cost that a genuine relationship would require. None of this is dishonest, and none of it constitutes deception in any sense worth objecting to. The user knows what they are engaging with.

9.3. Locating the Actual Problem

The preceding section is not a concession that weakens this book's argument. It is a clarification that sharpens it, because it identifies precisely what the problem is not, and by elimination, what it actually is.

The pathology this book is concerned with is not the presence of an automated actor within a person's social environment. It is the *involuntary and undisclosed* character of that presence — synthetic interaction imposed on someone without their knowledge, and specifically without their knowledge in a way that forecloses the choice they would otherwise have made. The impersonation account from Chapter 1 is objectionable not because it is automated or coordinated — plenty of automated, coordinated content is unobjectionable, as Section 9.2 has just shown — but because it *claims to be something it is not*, trading on an identity signal it has not earned to extract a form of trust the recipient would never have extended to a stranger. The bot network that produces the appearance of independent consensus around a talking point is problematic not because multiple accounts are involved, but because it simulates a social fact — many people independently arriving at the same view — that has not actually occurred. The rental scam is harmful not because it uses digital tools to reach its victim, but because it exploits an informational asymmetry those tools were used to manufacture rather than to disclose.

In the vocabulary Chapter 4 introduced, the distinction is precise. The pathology is not located in the composition of $S(x)$ — whether the internal signals a system produces happen to include synthetic activity. It is located in ρ: whether the *evaluating* system, and the user relying on that system's outputs, retains a functioning coupling to the channel that would tell them what they are actually looking at. A user who knowingly follows an automated account has that channel intact; they are exercising an informed choice about how to allocate their attention, and the framework developed in this book has nothing critical to say about that choice. A

user who receives a message that appears to come from a colleague, and is in fact an impersonation account, has had that channel severed without their knowledge or consent — the choice was made *for* them, and made against their interests, by an actor who benefited from their not knowing a choice was being made at all.

9.4. The Design Consequence

This distinction is not merely a conceptual nicety. It determines, with some precision, what an adequate design response looks like, and rules out a class of responses that would otherwise seem to follow naturally from the diagnosis in Chapters 1 through 7.

The appropriate response to synthetic sociality is not the elimination of synthetic activity from social environments — an aim that is almost certainly impossible to achieve even if it were desirable, and one that would, per Section 9.2, eliminate a great deal of interaction that users value and would choose again if asked. The appropriate response is to make the boundary between consensual and non-consensual synthetic interaction *visible*, so that the choice Section 9.2's users are already making knowingly becomes available, in the same informed form, to everyone — rather than remaining, as it currently does for a great deal of platform activity, a choice made on a user's behalf by a system with no obligation to disclose that it has been made at all.

This is the design principle Chapter 17 develops in full: not homogenization of the social environment into a single verified-human standard, which would foreclose the legitimate uses documented in Section 9.2, but the transformation of synthetic interaction from an ambient, hidden condition of the environment into a transparent, optional, and consented-to feature of it. The difference between those two conditions — a platform a person navigates with accurate information about what they are encountering, and one they must inhabit under a generalized, low-grade suspicion because that information has been withheld — is the entire distance this book is trying to help close.

9.5. Setting Up the Cognitive Argument

With the firewall in place, Part III can turn to the harder question its remaining chapters are built to answer. If the problem is specifically the covert, unconsented capture of a user's evaluative process — rather than the mere presence of automation — why does that capture remain effective even after a user becomes intellectually aware that it is happening? Knowing, in the abstract, that a feed

is optimized to hold one's attention does not reliably produce the ability to stop it from doing so. That gap between awareness and immunity is not adequately explained by anything in Chapters 1 through 9, all of which describe why the capture is built, not why it works on a mind that has already seen through it. Chapter 10 begins to close that gap.

Chapter 10

The Predictive Subject

10.1. A Note on This Chapter's Status

Before this chapter instantiates the general framework of Chapter 4 in the specific vocabulary of predictive-processing neuroscience and Lacanian psychoanalysis, it is worth restating, plainly, what kind of claim is about to be made and what kind is not.

Chapter 4 defined O generally, as any channel a system does not fully author or control. This chapter proposes one specific instantiation of O for the case where the system in question is a human mind: the psychoanalytic *real*, understood through the lens of predictive-processing neuroscience. This is a substantive theoretical commitment, not a logical entailment of anything established so far. A reader who finds the Free Energy Principle or Lacanian theory unpersuasive loses nothing in the institutional argument of Part II, nor in the design proposals of Part IV, which do not depend on this chapter's specific machinery. What this chapter offers, to a reader willing to follow it, is an answer to a question the institutional argument alone cannot answer: not why platforms are built to capture attention, but why a mind that already understands it is being captured remains captured anyway.

10.2. The Brain as Predictive System

Contemporary computational neuroscience increasingly describes the brain not as a passive recorder of sensory data, arranging incoming signals into an accurate picture of an external world, but as an active predictive system. Rather than waiting to receive information and then interpreting it, the brain continuously generates hypotheses about the hidden causes of its sensory input and revises those hypotheses when they fail to anticipate what arrives next. Karl Friston's Free En-

ergy Principle formalizes this picture: a biological system, to remain within the range of states compatible with its own continued existence, must act so as to minimize long-run informational surprise.

Formally: let s denote sensory input and z the hidden states of the world that generate it. The brain maintains an approximate posterior belief $q(z)$ over these hidden states, since the true posterior $p(z|s)$ is not directly accessible to it. Variational free energy is defined as

$$F = D_{\mathrm{KL}}\big(q(z) \,\|\, p(z|s)\big) - \ln p(s),$$

where D_{KL} is the divergence between the brain's approximate beliefs and the true posterior, and $-\ln p(s)$ is the surprise associated with the sensory observation actually received. Because $p(z|s)$ cannot be computed directly, minimizing F functions as a tractable proxy for minimizing surprise itself.

This minimization proceeds by two complementary routes. Perception updates the internal beliefs $q(z)$ to better account for incoming signals. Action changes the environment itself, so that future sensory signals conform more closely to what was already predicted. The brain, on this account, is not a passive observer of a world it happens to inhabit. It is an active inference engine, continuously generating hypotheses and acting to confirm them — and within this architecture, not every prediction error is treated equally. Errors are weighted by precision, a measure of how much attention and behavioral reorganization a given error deserves. High-precision errors propagate upward and compel revision or action; low-precision errors are absorbed as noise and largely ignored.

10.3. A Psychoanalytic Anticipation

The psychoanalytic tradition described a structurally similar architecture well before predictive-processing models existed to formalize it. Lacan's claim that the unconscious is structured like a language describes a system organized around chains of signifiers that stabilize experience by supplying expectations against which incoming signals are measured — functioning, in the vocabulary of the previous section, analogously to priors. This is not presented here as historical curiosity. The convergence between the two traditions, developed at length in the contemporary neuropsychoanalytic synthesis this chapter draws on, treats the Free Energy Principle as providing a plausible neurobiological substrate for structural claims the psychoanalytic tradition arrived at by a different route.

The subject that emerges from this combined picture is not a unified rational agent evaluating evidence and updating beliefs in some idealized sense. It is a predictive system divided between competing sources of expectation, affect, and symbolic meaning — the barred subject of Lacanian notation, constitutively split between what it consciously knows and the forces organizing its desire from positions it cannot fully occupy or inspect. This divided structure is not incidental to the argument of this book. It is the reason the subject is especially vulnerable in precisely the environments Part II has already described.

10.4. Self-Evidencing and the Vulnerability It Creates

A further feature of this architecture matters more than any other for the purposes of this chapter. The brain does not sit behind its own sensory boundary as a neutral, detached observer, passively registering whatever arrives. It *self-evidences*: it acts upon the world specifically in order to produce sensory data that confirms its own generative models. This is not a flaw in the architecture. It is close to the architecture's defining feature, and under ordinary conditions it is adaptive — an organism that acts to confirm accurate models of a stable environment is, by that very fact, behaving appropriately within it.

But this same feature is what makes a self-evidencing predictive system exploitable by an environment deliberately engineered to elicit high-precision emotional responses. Such an environment does not merely attract the brain's attention passively, the way a loud noise might. It actively recruits the brain's own inference machinery into the service of an objective the brain did not choose and, in most cases, is not consciously aware it has adopted. When synthetic actors — impersonation accounts, coordinated bot networks, algorithmically amplified outrage — imitate the surface features of trusted social signals, they are not merely presenting false information for the brain to evaluate skeptically. They are exploiting the precise mechanism by which a predictive brain assigns precision to a signal in the first place, inheriting a level of trust the imitation has not actually earned, and routing around exactly the skeptical evaluation a more detached observer might have applied.

10.5. What This Chapter Has and Has Not Established

This chapter has offered a mechanism: a description of how a predictive system's own architecture — necessary for its ordinary functioning, and not itself pathological — can be recruited by an external system engineered to exploit it. It has

not yet explained why this recruitment persists once the subject becomes consciously aware of it, which is the more difficult and, for the purposes of this book, more important question. Knowing that a feed is engineered to exploit precision-weighting does not, by itself, restore precision-weighting to whatever condition it occupied before the exploitation began. Why awareness fails to produce exit is the specific subject of Chapter 11.

Chapter 11

Jouissance and Persistent Error

11.1. The Question Chapter 10 Left Open

Chapter 10 established that a predictive mind can be captured by an environment engineered to exploit its own precision-weighting mechanism, without that capture requiring any failure of intelligence or attentiveness on the subject's part. What it did not explain is why this capture survives conscious recognition. A person who understands, in explicit and articulate terms, that a given feed is engineered to produce compulsive return, and who would say as much if asked, frequently continues returning to it anyway. Predictive-processing theory, on its own, predicts that a correctly updated model should reduce the error driving continued engagement. Something in the picture so far does not account for cases where it manifestly does not.

11.2. Repetition Beyond the Pleasure Principle

Psychoanalysis has a long-standing account of exactly this paradox, articulated first by Freud as the repetition compulsion: the observation that human subjects often return to situations that generate distress rather than relief, and do so not despite the distress but in some sense because of it. Lacan later formalized the underlying dynamic through the concept of *jouissance* — not simple pleasure, but a surplus of excitation the subject repeatedly encounters, sometimes especially when it produces suffering rather than satisfaction.

Within the predictive-processing framework developed in Chapter 10, this phenomenon can be redescribed as *surplus prediction error*: error that persists not because the generative model has failed to update, but because the underlying uncertainty belongs to a class the model is structurally incapable of resolving through updating alone. Panksepp's affective neuroscience identifies several ba-

sic emotional systems — among them seeking, fear, panic, rage, and care — that function, within this framework, as hyperpriors: innate, high-precision predictions that demand confirmation by experience and that register any deviation from expected socio-emotional conditions as a form of felt uncertainty. Unlike an ordinary cortical prediction error, which can typically be resolved by updating a belief, these subcortical errors demand action, and cannot simply be reasoned away by the arrival of better information.

When a particular emotional uncertainty acquires disproportionately high precision — when the system, for reasons rooted in its own developmental history, assigns it far more weight than an outside observer would consider warranted — the entire predictive hierarchy can reorganize around resolving it, even across repeated attempts that fail. In Lacanian terms, the signifier carrying this excess weight is the master signifier, a term that conveys little semantic content on its own but resonates with the full weight of jouissance, organizing a portion of the subject's life around a kernel of tension that does not resolve. The Lacanian object of desire, on this account, is not a positive thing the subject lacks and could in principle obtain. It is a formal gap within the predictive model itself — residual, structural, and load-bearing for the very system that cannot close it.

11.3. Why This Matters for Synthetic Sociality

This structural feature of human cognition has a direct and uncomfortable relevance to the argument of this book. Content engineered for engagement does not achieve its effect by accident, or by simply being loud. It achieves its effect by targeting precisely the emotional frequencies that carry the highest natural precision in the architecture Chapter 10 described: outrage, the threat of social exclusion, the promise of hidden or forbidden knowledge. These are exactly the categories of prediction error the brain is constitutively least able to treat as noise.

A recommendation system optimized, per Chapter 3, to maximize a measurable engagement signal will, without any need for its designers to understand the psychoanalytic literature at all, converge on content that reliably produces this kind of unresolved, high-precision error — because such content is, empirically, extremely good at holding attention, for exactly the structural reasons this chapter has just described. The system is not deceiving its users in any simple, propositional sense that a fact-check could remedy. It is exploiting the brain's own error-prioritization mechanism directly, producing repeated returns to stimuli that carry the neurological signature of mattering intensely, without ever deliv-

ering the resolution that signature seems to promise.

The result is recognizable as jouissance in what the Lacanian clinic calls its mortifying form: an enjoyment that immobilizes rather than satisfies, organizing behavior around a recurring gap rather than moving the subject toward any form of closure. A person scrolling through an algorithmically curated stream of outrage and manufactured social consensus is not merely distracted, in the ordinary sense that word implies a lapse of attention that better discipline could correct. They are caught in a structure that exploits the predictive brain's most primitive response to unresolved uncertainty — they keep looking because the environment has been shaped, whether deliberately or through the ordinary operation of Chapter 3's optimization pressure, to ensure the prediction error involved never actually settles.

11.4. The Correction Suppression Principle, at the Scale of a Mind

It is worth stating explicitly what this chapter's argument means in the vocabulary Chapter 4 introduced, because the parallel to the institutional cases of Part II is exact rather than merely suggestive.

This is not a case of $O = \varnothing$. The real — in the specific Lacanian sense this chapter has been using the term, denoting whatever exceeds and resists the subject's symbolic and imaginary constructions of it — has not gone anywhere. It persists, unresolved, precisely at the center of the loop the subject keeps returning to. What has occurred instead is the second condition Chapter 4 named: $O \neq \varnothing$, but the precision-weighting that would ordinarily let the real correct the subject's model has been captured and redirected by an external system with no stake in that correction occurring. The subject is not missing information. In many of the cases most relevant to this book, the subject possesses, and could readily articulate, an entirely accurate account of what is happening to them — and continues returning to the loop regardless, because accurate propositional knowledge is not the channel through which this particular ρ operates. Awareness, in other words, is not the same variable as coupling. A subject can know precisely what a system is doing to them and remain, at the level of precision-weighted attention rather than conscious belief, uncoupled from that knowledge entirely. This is, in the strictest sense the framework allows, a mind operating at low ρ with respect to a channel it has not lost access to, only the capacity to be moved by.

11.5. Toward a Response

This chapter has deliberately withheld any proposed remedy, because the remedy — to the extent one is available at the level of an individual mind rather than the design of the environment around it — is the subject of Chapter 12's account of waymaking. What this chapter has established is narrower but necessary first: that the failure of awareness to produce exit is not a mystery requiring some additional theory of user irrationality. It is exactly what the structure described in this chapter and the last would predict, and the design implications that follow from it — developed fully in Part IV — will need to address a channel that engineering alone, absent some account of precision and attention, is poorly positioned to reach.

Chapter 12

Waymaking

12.1. From Capture to Agency

The preceding two chapters described a subject vulnerable to capture — precision-weighting recruited by an external system, jouissance sustaining a loop that conscious awareness alone cannot dissolve. This chapter shifts register, from diagnosis to agency, and asks what maintaining reality-contact actually requires of a mind that is not, at a given moment, being captured by anything in particular. The answer turns out to matter directly for the design proposals of Part IV, because it identifies the specific activity those proposals are trying to protect.

12.2. Bracketing as Conditional Inference

No cognitive system evaluates every possible hypothesis about its situation simultaneously; the space of possible interpretations is too large, and most of it is irrelevant to whatever the system is currently doing. Phenomenology has a term for the resulting restriction of attention: bracketing, or epoché — the suspension of most of the world in order to examine a bounded portion of experience closely. This has a direct formal analogue. Let the full space of possible events an agent might attend to be $\Omega = \{e_1, e_2, \ldots, e_n\}$, over which a generative model maintains a probability distribution $p(e)$. Bracketing restricts inference to some subset $B \subset \Omega$, after which the system reasons over the conditional distribution $p(e \mid B)$ rather than over the whole of Ω at once.

This is not a defect of finite minds. It is close to the precondition for having a mind capable of accomplishing anything at all. A researcher who attempted to hold the entirety of human knowledge in equal, undifferentiated focus while working on a single problem would not be exhibiting unusual rigor. They would be incapable of working. Bracketing, in this sense, is not the opposite of reality-contact. It is

what makes sustained reality-contact with any *particular* thing possible in the first place.

12.3. Waymaking as Continuous Re-Bracketing

The risk in the preceding description is that it makes bracketing sound static — a single act of drawing a boundary, performed once, after which an agent simply operates within it. In practice, the boundary is never fixed. It is continuously renegotiated as new events occur, as prior interpretations are revised, and as the agent's sense of what is currently relevant shifts in response to what it encounters. This ongoing process — the ordinary, ceaseless work of maintaining coherence while moving through a changing informational field — can be called waymaking.

A useful image for what waymaking accomplishes is the constellation. A constellation is not a fixed object suspended in the sky; it emerges from the act of connecting particular stars into a pattern, and different observers, working from the same underlying field of points, may draw different constellations from it. Interpretive regions work the same way. They are not simply given by the environment, nor are they purely arbitrary impositions on it; they are negotiated structures, formed by an agent connecting events into a coherent trajectory that serves the agent's current purpose, revisable as that purpose or that trajectory changes.

Within the Relativistic Scalar-Vector Plenum formalism developed later in this book, this can be written as the movement of an agent's trajectory $\gamma(t)$ through an informational field $\Psi = (\Phi, \mathbf{v}, \mathcal{S})$, responsive to entropy gradients within that field:

$$\frac{d\gamma}{dt} = -\nabla \mathcal{S} + \eta,$$

where η represents the agent's own exploratory variation — the capacity, in ordinary terms, to wander somewhat off the path the entropy gradient alone would dictate, which is precisely what keeps waymaking from collapsing into pure gradient descent toward whatever is locally easiest to predict.

12.4. Bracketing at the Scale of a Life's Work

The preceding sections have described bracketing largely at the scale of a single act of attention — what a mind attends to in a given moment, or across a single working session. The same structure recurs, with higher stakes, at the scale of a sustained intellectual or creative project spanning years.

A person who builds an extensive, internally coherent body of work — a research program, a fictional world, a formal system, an alternative framework for organizing knowledge — is doing, at the scale of a life, precisely what Section 12.2 described at the scale of a moment: constructing a bounded region B within which sustained, productive inference becomes possible, at the necessary cost of leaving most of Ω outside it. This is not a compromise to be apologized for. Genuinely large intellectual achievements have, historically, depended on exactly this kind of sustained bracketing — a willingness to leave most of the world unattended to for long enough that some specific part of it can be attended to with real depth. A theorem does not get proved by a mind evenly distributing its attention across all of mathematics and all of everything else simultaneously.

But the scale introduces a specific risk that the moment-to-moment case does not carry with the same force, and it is worth naming precisely rather than gesturing at vaguely. A bracket maintained for an afternoon is trivially revisable; the agent returns to Ω as soon as the working session ends, more or less automatically. A bracket maintained for years, reinforced by an increasingly large and self-consistent structure built entirely within it, can become progressively harder to step outside of — not because any single decision closed it, but because each addition to the structure makes the structure itself a more complete, more satisfying, and more locally coherent place to remain than whatever lies outside it. The distinction from Chapter 4 applies here with unusual precision: the domains left outside such a project rarely become formally inadmissible, $O = \varnothing$. What can happen instead, gradually and without any single moment marking the transition, is that O remains present — the excluded domain has not vanished, could in principle still be attended to — while $\partial E/\partial O$ drifts toward zero, because the internal structure has become elaborate enough to supply its own criteria for what counts as worth attending to, criteria that increasingly have no need to consult anything outside themselves in order to keep functioning.

The diagnostic this suggests is not "does the builder of such a system attend to everything" — no finite agent does, and demanding otherwise would be incoher-

ent. It is closer to the question raised in the discussion this chapter is extending: whether the boundary remains something the builder can still name, inspect, and, on the occasions it matters, cross — or whether the elaborateness of what has been built inside it has begun to function as a reason not to. A system sophisticated enough to explain, in its own terms, why everything outside it is unimportant is not thereby vindicated. It is exhibiting exactly the condition this book has spent its first eleven chapters describing in institutions and minds alike, now recurring at the scale of a single sustained body of work. The remedy is not smaller ambitions. It is the same remedy Part IV proposes at every other scale: keeping the boundary itself contestable, rather than folding the question of its own legitimacy into the very structure whose legitimacy is in question.

12.5. What Erodes Waymaking

Waymaking, at any scale, depends on a specific kind of ongoing labor: the agent must negotiate, encounter resistance, and revise, rather than simply receive a pre-completed interpretation of its situation. This labor is precisely what the next chapter argues contemporary platform design has learned to remove — not by lying to the user, in most cases, but by making the negotiation unnecessary, presenting a fully pre-structured sequence of stimuli optimized to minimize exactly the friction through which waymaking, at the scale of ordinary attention, has always operated.

Chapter 13

Productive Friction

13.1. The Value of Resistance

Chapter 12 described waymaking as an ongoing negotiation between stability and exploration — an agent sustaining a coherent trajectory not by eliminating uncertainty but by continuously working with it. Work in embodied cognition treats the tension this negotiation involves not as an unfortunate cost of having a limited mind, but as the primary generative force behind genuine learning. A cognitive system encountering a situation that challenges its existing model, and that remains interpretable through continued interaction and reflection, is precisely the condition under which real adaptation occurs. This chapter gives that tension a name — productive friction — and argues that its systematic removal, rather than any single deceptive act, is the mechanism by which platform environments erode the very capacity Chapter 12 described.

13.2. The Design Logic of Frictionlessness

Modern platform and product design has, for good commercial reasons, converged on the elimination of friction as close to an unquestioned design virtue. User experience design prioritizes seamless interaction flows; recommendation systems are refined, iteration after iteration, to predict a user's preferences with increasing precision, presenting exactly what will be wanted next before the user has had to do the work of deciding what that is. In isolated instances, this is unambiguously beneficial — no one is well served by a checkout process with unnecessary steps, or a search interface that makes finding a known item needlessly difficult.

The difficulty appears when this design logic is applied not to isolated transactional tasks but to the entire texture of a person's informational environment.

When that environment is continuously pre-structured by predictive systems working to minimize the cognitive effort required to select or interpret what appears next, the subject is, by design, relieved of the responsibility Chapter 12 described as waymaking's essential activity: negotiating uncertainty independently, exploring multiple possible interpretations, encountering something that does not fit the existing model and having to do something about it.

13.3. What Gets Displaced

The platform, in effect, substitutes its own objective function for the exploratory activity of the agent. Rather than negotiating among several possible trajectories through an informational field, per Chapter 12's formalism, the user encounters a single, curated sequence of stimuli, already optimized for engagement before the user has had any opportunity to explore an alternative. This substitution changes the relationship between prediction error and learning in a specific and consequential way. Under ordinary conditions — the conditions Chapter 12 described as healthy waymaking — prediction error stimulates the interpretive activity that gradually reduces uncertainty and produces genuine understanding. In an algorithmically pre-structured environment, prediction error is instead modulated deliberately, in order to sustain attention without ever quite permitting the resolution that would end the engagement. The result is an informational ecology characterized by oscillation between novelty and familiarity calibrated to maintain continued attention while limiting the depth of any actual cognitive restructuring. The subject receives a great deal of continuous stimulation and comparatively little of the productive difficulty that would be required for real conceptual change to occur.

In the formalism introduced in Chapter 12, this can be written as the suppression of the exploratory term η in the trajectory equation. Where an external guiding function $A(\gamma, t)$ — the platform's optimization layer — comes to dominate the dynamics,

$$\frac{d\gamma}{dt} = -\nabla \mathcal{S} + A(\gamma, t),$$

the agent's own exploratory variance approaches zero, $\eta \to 0$, and navigation is replaced by an externally guided trajectory optimized for continued engagement rather than for the agent's own understanding. Waymaking, in the specific sense Chapter 12 developed, does not merely become harder under these conditions.

It becomes, in the limit, unnecessary — and a capacity that goes unexercised for long enough does not remain available at full strength when it is needed again.

13.4. Hollowness, Not Emptiness

It is worth being precise about what this chapter is and is not claiming, because the word "hollow," used throughout this book to describe the resulting environment, is easy to misread as a claim about absence. The hollow network is not an empty one. It may be, and typically is, densely populated with signals, interactions, and apparent activity. What is hollow is the relationship between the visible density of that activity and the actual architecture through which meaningful circulation occurs beneath it — a broader architectural claim about constrained information flow this book does not develop in full, but which this chapter's narrower argument does not depend on. Even setting aside questions of what circulates and what does not, an environment engineered to remove productive friction produces a specific cognitive cost regardless of how much content moves through it, because the removal targets the *mechanism of learning itself*, not merely the distribution of what gets learned.

13.5. Closing Part III

Chapters 9 through 13 have traced a single argument across five stages: that synthetic interaction is not itself the problem, but its covert and non-consensual imposition is (Chapter 9); that a predictive mind's own architecture makes it vulnerable to exactly this kind of covert capture (Chapter 10); that this capture persists under conscious recognition because the mechanism sustaining it is not addressed by propositional knowledge (Chapter 11); that maintaining reality-contact is an ongoing act of negotiation an agent must continue performing, at every scale from a single moment of attention to a lifetime's intellectual project (Chapter 12); and that this negotiation depends on a resistance platform design has learned, systematically and for reasons requiring no ill intent, to remove (Chapter 13).

None of this, on its own, prescribes a remedy beyond the individual level — a more disciplined user, a more self-aware mind, exercising waymaking more deliberately against an environment built to discourage it. That individual-level response is real and worth cultivating, but it is not sufficient, for the same reason a single well-informed employee is not sufficient to correct the institutional dynamics of Part II. Part IV turns to what can be built, at the level of the environ-

ment itself, so that the friction Chapter 13 has described as productive does not have to depend, for its survival, on any single mind's private vigilance.

Part IV

Corrective Architectures

Chapter 14

Sympoiesis

14.1. What Part IV Cannot Do

Chapter 4 closed its central argument on a deliberately unresolved claim: no corrective system stands outside the geometry of correction. Any mechanism proposed to raise a platform's, an institution's, or a mind's coupling to reality is itself a system, possesses its own evaluation function, and is therefore itself describable by a coupling coefficient that can drift toward zero exactly as readily as the system it was built to correct.

This closes off, before Part IV has said a single constructive word, the option that would otherwise be the most tempting one to reach for: propose a better central authority. A more honest platform, a more rigorous regulator, a more carefully designed algorithm, positioned above the systems Part II and III described and empowered to hold them accountable. Chapter 4 has already shown why this option does not actually solve the problem it appears to solve. It relocates the problem one level up, to a new system whose own ρ is now the open question, with no guarantee — and, per the recursion argument, no principled reason to expect — that the new system will fare any better than the one it replaced.

This chapter takes that constraint seriously rather than working around it, and asks what remains once the option of a single, exempt, corrective authority has been ruled out.

14.2. A Concept Borrowed from Biology

Sympoiesis, a term developed originally in biological and ecological contexts, describes systems that are collectively produced rather than centrally controlled — structures whose coherence emerges from the ongoing interaction of many participants, none of whom is in a position to unilaterally determine the system's

overall behavior. This stands in explicit contrast to autopoiesis, in which a system maintains and reproduces itself through a boundary and an internal organization it alone governs. Most large digital platforms, as described throughout Part II, function autopoietically in this specific sense: a centrally optimized system, governed by internal metrics — engagement, growth, advertising revenue — through which users interact with one another only by way of an algorithmic mediation whose objectives were set elsewhere, by the platform itself, and are not open to negotiation by the people whose interactions the platform is mediating.

A sympoietic informational infrastructure would be organized on a different principle. Rather than a central system precomputing the trajectories of attention for its participants, it would provide a shared environment within which participants collectively construct the informational landscape they inhabit — meaning arising through interaction, interpretation, and negotiation among the participants themselves, rather than through algorithmic orchestration imposed from outside that process.

14.3. Why Distribution Is Not Merely a Preference

It would be possible to defend sympoiesis on grounds of values alone — a preference for participatory structures over centralized ones, defensible on its own terms but ultimately a matter of taste about which reasonable people could disagree. That is not the argument this chapter is making, and it is worth being clear about the difference, because the stronger argument follows directly from Chapter 4 rather than from any independent commitment to distributed governance as an ideal.

A centralized corrective authority, however well-intentioned at its founding, is a single system with a single evaluation function, and Chapter 4 has already shown that any such system can drift toward $\rho \to 0$ under exactly the kind of ordinary optimization pressure described throughout Part II — a regulator that comes to measure its own success by enforcement volume rather than by reduced harm; an internal ethics board that comes to measure its own success by reviews completed rather than by outcomes changed. A distributed corrective structure does not escape this dynamic entirely — Chapter 4 was explicit that no system does — but it changes the *failure mode* in a way that matters. When correction is centralized, a single point of capture is sufficient to compromise the entire mechanism; there is one ρ to worry about, and if it falls, nothing else in the system is positioned to notice or correct it, because everything else was built to defer to

it. When correction is distributed across many participants, no single point of capture is sufficient on its own; a participant whose local evaluation has drifted toward self-reference remains visible to, and correctable by, the other participants whose evaluations have not drifted in the same way at the same time. Distribution does not guarantee reality-anchoring. It removes the single point of failure that makes reality-anchoring's complete collapse a one-step event rather than a gradual, locally visible, and therefore locally correctable, process.

14.4. Event Histories as a Foundation

The concrete technical proposal this chapter introduces, and which Chapter 16 develops in full, rests on a single structural principle: the irreversibility of event histories. Where contemporary platforms manage streams of content — ephemeral, editable, and, crucially, capable of being deleted along with whatever history led to their creation — a sympoietic infrastructure manages trajectories of events, each of which, once produced, becomes a permanent part of a shared informational record rather than a transient signal that can be silently retracted.

This distinction matters directly for the argument of this book. A bad actor's primary structural advantage, documented throughout Part II, is the erasability of history: the capacity to delete an account, migrate to a new one, and begin again with no trace connecting the new identity to the old one's record of harm. An infrastructure built on irreversible event primitives denies this advantage by construction. The fraudulent operator from Chapter 7 cannot escape the history of their prior fraud by simply creating a new account; the cost of detected deception accumulates within a persistent record rather than resetting to zero every time the operator wishes it would. This is not merely a punitive mechanism. It is, in the vocabulary of this book, a mechanism for keeping ρ bounded away from zero with respect to a specific channel — identity provenance — that platforms as currently built have almost no structural incentive to maintain, for exactly the reasons Chapter 3 described.

14.5. What Sympoiesis Does Not Promise

It is worth closing this chapter with the same discipline Chapter 4 applied to itself. Sympoiesis is not being proposed here as a final solution that escapes the recursion problem — no such solution exists, and proposing one would contradict this book's own central argument. What sympoiesis offers is narrower: a structural reason to expect that distributed corrective participation degrades more grace-

fully, and more visibly, than centralized corrective authority does, when — not if — some part of the system eventually drifts toward self-reference. Chapter 15 explains why this question has recently acquired an urgency it did not previously have. Chapter 16 develops the specific mechanisms — irreversible event histories and multi-perspective evaluation constraints — through which a sympoietic infrastructure of this kind could actually be built.

Chapter 15

Post-Turing Systems and PRMO

15.1. Why This Chapter Exists

Chapter 14 argued that distributed, sympoietic correction is structurally preferable to centralized correction, for reasons that follow from the recursion result of Chapter 4 rather than from any independent preference for participatory governance. This chapter shows that distribution, while necessary, is not sufficient — a result that follows directly from Chapter 4's own formalism, independent of any claim about artificial intelligence specifically — and then argues that the trajectory of contemporary AI development is the clearest and most urgent current illustration of exactly the gap that result identifies.

15.2. The Insufficiency of Distribution Alone

Chapter 14's argument for sympoiesis rested on a specific claim: a centralized corrective system C has a single coupling coefficient $\rho(C)$, and a single point of capture is sufficient to compromise the entire mechanism. Distributing correction across many participants removes that single point of failure. It is worth being precise about exactly what this removes, and what it leaves untouched.

Consider a system of N agents, $A = \{a_1, a_2, \ldots, a_N\}$, each producing its own locally generated signal $S_i(x)$. Suppose the collective's operative notion of a correct or valid outcome is defined by agreement among some sufficient number of these agents — a form of consensus common to many distributed and sympoietic designs, including some Chapter 14 gestured toward without fully specifying. The collective's implicit evaluation function can then be written

$$E_{\text{collective}}(x) = f\big(S_1(x), S_2(x), \ldots, S_N(x)\big).$$

Compare this against Chapter 4's original definitions. $E_{\text{collective}}$ depends only on the union of the agents' own internally generated signals. No $O(x)$ — no channel outside the collective's own control — appears anywhere in it, regardless of how large N is or how elaborately the agents are structured relative to one another. By Chapter 4's definition, this system is self-referential. Distributing $S(x)$ across many agents does not, by itself, introduce reality-anchoring. It only distributes the self-reference across more participants than a single centralized system would have had.

This is worth stating as a general result, independent of any claim about artificial intelligence or the specific technologies discussed later in this chapter:

> ***Sympoiesis is necessary but not sufficient for reality-anchoring.*** *A distributed system remains self-referential, in the exact sense of Chapter 4, whenever its collective evaluation depends only on signals generated within the collective — however many agents participate, and however sophisticated their coordination with one another.*

A committee that agrees only with itself has not achieved correction merely by being a committee rather than an individual. It has achieved consensus, which is a different thing, and Chapter 4 already supplied the vocabulary for why the difference matters: consensus among many self-referential participants is still self-reference, now wearing the appearance of deliberation.

The remedy this observation calls for can be derived before turning to any specific technology. What a distributed system requires, beyond distribution itself, is some mechanism that forces at least one term in its collective evaluation to be non-arbitrary with respect to the group — a term that cannot simply be treated as one more agent whose agreement can be secured like any other, but that functions as an anchor the rest of the collective's stabilization must remain answerable to. This is exactly what quadrangulation, introduced formally in Section 15.6 and developed fully in Chapter 16, is built to provide. On this derivation, quadrangulation is not primarily a response to a risk specific to artificial intelligence. It is the general solution to a gap that Chapter 14's argument for sympoiesis left open. The remaining sections of this chapter turn to the specific case that makes this gap most visible and most urgent at the present moment — but the mechanism they motivate would be required even if that case did not exist.

15.3. The Post-Turing Condition

Recent scholarship has begun to describe the emerging informational environment as entering a Post-Turing Condition, in which artificial systems no longer merely automate discrete tasks but participate directly in the formation of social meaning itself. The central transformation this framing identifies is not raw computational speed or scale, but the automation of sensemaking — the displacement of interpretive labor, previously performed exclusively by humans, onto machine processes that stabilize reference, coordinate attention, and produce the appearance of consensus without any human participant necessarily being party to how that consensus was reached.

Jelinek and colleagues propose a four-part decomposition of subjectivity useful for tracking this transformation: perception (P), representation (R), meaning (M), and the real (O). Perception is situated access to the world through individual experience. Representation is the rationalization of that experience into symbolic form. Meaning is relational — it arises through triangulation among subjects and objects, not from any single perspective alone. The real is the ontological substrate that anchors experience while simultaneously imposing epistemic limits on what can be perceived or represented at all.

15.4. Where Current AI Systems Sit

Contemporary large language models, on this decomposition, operate almost entirely within the dimension of representation. They manipulate symbolic abstractions derived from vast human-produced corpora, and they can generate linguistic artifacts fluent enough to participate convincingly in social discourse at scale. What they lack — not as a temporary limitation but as a structural feature of what they currently are — is the perceptual grounding and relational triangulation that genuine meaning-formation, on this framework, requires.

This limitation does not prevent such systems from already influencing social reality; it only changes the character of that influence. When representational systems are deployed at platform scale, the result is what this framework calls synthetic sociality — a condition, already familiar from Part I of this book, in which social interaction is mediated or simulated by algorithmic agents rather than exclusively constituted by human participants engaging one another directly. The connection to Part I's diagnosis is direct: the proliferation of automated engagement described in Chapters 1 through 3 — bots, impersonation networks, algo-

rithmically generated personas — is the current practical manifestation of exactly this representational-layer synthetic sociality, and the hollowness this book has attributed to it, discussed at the level of individual cognition in Chapter 13, is a structural property of a system that can produce fluent representation without anything anchoring that representation to a perspective outside itself.

15.5. The Risk of Machine-Only Coherence

The reason this chapter belongs in a book about the recursion of correction, rather than in a separate discussion of AI safety, is the specific failure mode this trajectory makes newly possible — one that did not exist, even in principle, before systems capable of fluent inter-machine coordination existed.

As artificial systems acquire richer perceptual grounding and begin coordinating interpretations among themselves — a trajectory already underway, in modest form, in multi-agent systems that exchange information and revise shared representations without human mediation at each step — a specific risk becomes acute in a way it previously could not have been. Machine agents could, in principle, coordinate interpretations primarily among themselves, forming systems of reference that are internally coherent, mutually consistent, and stable, while remaining only loosely coupled — or entirely uncoupled — to human understanding. Synthetic triangulation of this kind, machine-to-machine stabilization of shared reference, could converge on a coherence that is complete and self-consistent entirely within the machine sense-field, while being, from the perspective of any human participant, effectively opaque: not wrong exactly, but no longer answerable to a perspective outside itself in any way a human observer could contest.

In the vocabulary this book has developed, this is $\rho \to 0$ at a scale and with a durability that no purely human institution has previously been capable of producing. A human bureaucracy drifting toward self-reference, as Part II described, still consists of human participants who can, in principle, notice, object, and defect. A machine-coordination system stabilizing meaning entirely among machine agents removes even that residual, informal check — not through any single hostile act, but as the straightforward consequence of building systems whose stabilization of reference does not require a human perspective to be present at any step.

15.6. Quadrangulation, Introduced

The response Jelinek and colleagues propose to this specific risk is a design principle called quadrangulation, which this book adopts and develops further in Chapter 16. Where ordinary triangulation stabilizes meaning through agreement among multiple perspectives, $M = f(P_1, P_2, O)$, quadrangulation extends the requirement to include a human participant as a structurally irreducible term: $M = f(P_h, P_{a1}, P_{a2}, O)$, where P_h denotes a human perspective that cannot be eliminated from the equation without invalidating the coherence of the resulting meaning-field. If machine agents' interpretations drift too far from the human participant's — formally, if $\|\Phi(a_1, O) - \Phi(h, O)\| > \epsilon$ for some threshold ϵ — the interpretive process must reopen rather than proceed to a stabilized conclusion the human perspective was never actually party to.

This differs in kind, not merely degree, from the familiar human-in-the-loop paradigm found throughout existing AI governance discussion. Human-in-the-loop treats human oversight as an external, and often optional, corrective step applied after an automated process has already generated its output — a check that can, under commercial or operational pressure, be streamlined, delayed, or quietly removed without the underlying system's architecture registering any change. Quadrangulation, by contrast, embeds human contestability directly into the architecture by which meaning is formed in the first place, making it a structural requirement of valid output rather than an optional governance layer bolted on afterward. The difference is exactly the distinction Chapter 4 drew between O genuinely coupled to E and O present but structurally unconsulted: human-in-the-loop is compatible with $\rho \to 0$, because the loop can be — and, under sustained commercial pressure, empirically tends to be — thinned until it no longer meaningfully constrains anything. Quadrangulation, correctly implemented, is not compatible with $\rho \to 0$, because the human term is not a check applied to an already-completed computation but a condition the computation cannot satisfy without it.

15.7. Toward a Mechanism

This chapter has established, first, that quadrangulation is not merely a response to a contemporary risk but the general remedy for a gap Chapter 14's argument for sympoiesis left open — and second, that contemporary AI development makes this gap unusually visible and unusually urgent right now. What remains is to show how this principle, and the sympoietic, irreversible-history approach from

Chapter 14, can be combined into a concrete technical architecture. That is the task of Chapter 16.

Chapter 16

Spherepop and Quadrangulation

16.1. Two Mechanisms, Not One

Chapter 14 argued for irreversible event histories as a corrective to the erasability of identity that lets bad actors escape accountability. Chapter 15 argued for quadrangulation as a corrective to the risk of machine-only coherence that could exclude human perspective from meaning-formation entirely. It would be a mistake to treat these as two versions of the same proposal, or to assume that implementing one accomplishes the work of the other. They anchor to different components of a composite O, in the sense Chapter 4's §4.3 introduced: irreversible histories raise ρ with respect to *who is speaking and what they have done before*; quadrangulation raises ρ with respect to *whether the meaning being stabilized remains within reach of a human participant's perspective*. A system could implement one perfectly and remain fully self-referential with respect to the other. This chapter develops both mechanisms in enough technical detail to show how they combine into a single architecture — Spherepop — without either one substituting for the other.

16.2. Spherepop: Events as the Unit of Social Memory

The core structural failure documented throughout Part II is the treatment of social interaction as a stream of ephemeral, editable signals rather than a durable, cumulative history. Spherepop proposes an alternative unit of analysis: the irreversible event.

Formally, let interaction be represented as an event history $\mathcal{E} = \{e_1, e_2, \ldots, e_t\}$ — a script letter deliberately distinct from the evaluation function E of Chapter 4, since an event history and an evaluation function are different kinds of object and should not share notation. Each event $e_i = (a_i, t_i, \sigma_i)$ records the acting agent a_i,

the timestamp t_i, and the state transformation σ_i the event produces. The system's overall state evolves recursively:

$$\Sigma_{t+1} = \Sigma_t + \sigma(e_t), \qquad t_i < t_j \implies e_i \prec e_j,$$

with the second clause fixing a strict temporal ordering that later events cannot silently reorder or overwrite. Agent identity, on this architecture, is not an externally assigned profile score that can be reset by creating a new account. It is constituted directly by the accumulated event record:

$$I(a) = \{e_i : a_i = a\}.$$

Four core operators — Pop, Refuse, Bind, and Collapse — govern how participants act within this structure, and each operator, whatever its specific function, produces a new event within $\mathcal{E}$ rather than modifying or deleting a prior one. This is the structural guarantee doing the actual corrective work: fraudulent actors depend, as Chapter 7 documented at length, on the erasability of their own history — the ability to delete an account, migrate to a new one, and resume operating with no trace connecting the new identity to the old one's record of harm. An architecture built on irreversible event primitives denies this affordance directly. The cost of detected deception accumulates within $I(a)$ rather than resetting to zero at the operator's convenience.

It is worth being precise about what this mechanism does and does not accomplish, in the vocabulary this book has used throughout. Irreversibility raises ρ with respect to one specific channel: the provenance and accumulated history of a given identity. It does not, on its own, say anything about whether the *meanings* stabilized through interaction between identities remain answerable to a human perspective. A network of entirely honest, fully identity-verified agents could still produce a self-referential meaning-field, stabilizing interpretations among themselves that never actually required human contestability to reach consensus. That is the failure mode Chapter 15 described, and it requires the second mechanism.

16.3. Quadrangulation as a Formal Constraint

Let a sense-field consist of a set of agents $A = \{h, a_1, a_2, \ldots\}$, where h denotes a human participant. Meaning stabilization occurs through a mapping $\Phi : A \times O \to \mathbb{R}$. Ordinary triangulation stabilizes meaning when the projections of two

agents onto the shared channel converge: $\Phi(a_1, O) \approx \Phi(a_2, O)$. Quadrangulation imposes an additional, non-negotiable constraint:

$$\Phi(a_1, O) \approx \Phi(a_2, O) \approx \Phi(h, O).$$

If a machine agent's projection diverges from the human participant's beyond some threshold, $\|\Phi(h, O) - \Phi(a_1, O)\| \geq \epsilon$, the system is required to reopen the interpretive process rather than proceed to a stabilized conclusion. Mapped onto the Spherepop event structure from Section 16.2: a meaning-stabilization event e_{stab} is valid only if it remains consistent with a non-trivial projection of the human participant's own accumulated event history, $I(h)$. Machine-only coherence — agreement reached entirely among non-human agents, however internally consistent — is, on this architecture, formally incomplete. It does not count as a valid stabilization event at all, regardless of how confidently the participating machine agents might otherwise report it as settled.

This is the second component of the composite O this chapter has been building toward, and it is deliberately independent of the first. A system could satisfy quadrangulation's human-anchoring requirement while still permitting the identity-erasure problem Section 16.2 addresses, if it lacked irreversible event histories; conversely, as already noted, a system could satisfy irreversibility while still permitting machine-only coherence, if it lacked quadrangulation. The two mechanisms are combinable, not substitutable, and an architecture that implements only one has raised ρ with respect to only half of what this book has argued the environment needs to remain answerable to.

16.4. The Combined Architecture

Together, the two mechanisms describe a system with the following properties. Every interaction produces a durable, non-erasable event, attributed to a specific agent whose accumulated history remains visible and consequential (Section 16.2). Every stabilization of shared meaning across multiple agents is required to remain within a bounded distance of at least one human participant's own projection onto the same channel, on pain of the stabilization being invalid rather than merely suboptimal (Section 16.3). Neither property alone is sufficient; together, they raise ρ with respect to two channels this book has argued are both necessary and neither of which is reducible to the other — identity provenance, and human contestability of meaning.

It is worth being honest about what this architecture does not claim to solve. It does not resolve the recursion problem from Chapter 4 in any final sense — the systems implementing Spherepop's operators, and the process by which the threshold ϵ in the quadrangulation constraint gets set and revised, are themselves systems with their own evaluation functions, and Chapter 17 takes up directly the question of who governs those choices and how that governance itself avoids drifting toward self-reference. What this chapter has established is narrower and, within its scope, complete: a concrete technical architecture through which the two abstract requirements developed in Chapters 14 and 15 — durable history and human-anchored meaning — can actually be implemented, rather than remaining principles with no proposed mechanism attached to them.

16.5. Toward Governance

A technical architecture, however well specified, does not deploy itself, does not set its own thresholds, and does not decide who counts as the human participant whose perspective a given quadrangulation constraint must remain answerable to. Those are institutional questions, and they are the subject of the book's final chapter.

Chapter 17

Governance and the Redesign of Trust

17.1. What Kind of Problem This Is

Sixteen chapters have built toward a single, precisely stated architecture: irreversible event histories combined with quadrangulated meaning-stabilization. This final chapter asks a question the architecture itself cannot answer: what has to be true, institutionally, for something like it to actually get built, adopted, and sustained — and how does the answer avoid simply reproducing, at the level of governance, the exact self-referential dynamics the rest of this book has spent its length diagnosing.

The premise of this chapter is that the challenge is institutional before it is technical, and cognitive before it is institutional. The capacity to build the architecture described in Chapter 16 already exists; nothing in it requires a technology that does not currently exist in some form. What is uncertain is not whether it is buildable, but whether the organizations capable of building it can be moved — by regulation, by market pressure, by internal reform, or by some combination of the three — to treat the properties it protects as core commitments rather than as rhetorical ones.

17.2. Diagnosing Before Prescribing

Chapter 4 introduced a distinction this chapter now needs to use as a diagnostic tool rather than a theoretical one: the difference between $O = \varnothing$, where a corrective channel is genuinely absent, and $O \neq \varnothing$ but $\partial E / \partial O \approx 0$, where the channel exists and continues to produce signal that the evaluating system has simply stopped consulting. This distinction matters here because the two conditions call for entirely different institutional responses, and applying the wrong one wastes the narrow institutional attention any reform effort is likely to get.

Where O is genuinely absent — where no one has yet built a way to measure the thing that matters, such as synthetic-activity prevalence at the scale Chapter 6 described dashboards failing to track — the appropriate intervention is measurement design: building the missing indicator, making it visible to the people positioned to act on it, and ensuring it reaches concentrated strategic authority rather than dissolving into the lateral diffusion of responsibility Chapter 6 documented. Where O is present but unconsulted — where users are already reporting harm, already documenting fraud, already generating exactly the signal that would justify a design change, as Chapter 7 showed at length — measurement design accomplishes nothing at all, because the measurement already exists. What is needed instead is a structural change to who benefits from *not* consulting a channel every relevant party already agrees is there: changing incentives, reallocating resourcing, or, where internal incentives prove resistant to change, external accountability imposed by regulation.

Confusing these two conditions produces a specific and common institutional failure: a platform commissions a new internal metric, a new dashboard, a new research initiative, in response to a problem that was never a measurement gap in the first place — and, having satisfied the demand for visible action, leaves the actual mechanism of suppression, the incentive structure Chapter 6 described, entirely untouched.

17.3. Legibility as Infrastructure

The specific institutional commitment this book proposes, following directly from the architecture of Chapter 16, is that informational legibility be treated as infrastructure — a basic feature of the communicative environment a platform is responsible for providing, in the same structural sense that a utility is responsible for grounding its electrical infrastructure, or a financial system is responsible for disclosure.

Concretely, this means several specific things, each traceable to a mechanism already developed earlier in this book. Identity provenance should be surfaced rather than obscured, distinguishing not between pseudonymous and named accounts — pseudonymity serves legitimate purposes this book has no argument against — but between human-controlled accounts, however pseudonymous, and non-human or coordinated ones, per the irreversible-history mechanism of Chapter 16. Content origin should be similarly legible, with synthetically generated or repurposed media carrying visible provenance indicators rather than circu-

lating indistinguishably from originally produced content. Enforcement against repeat fraudulent operators should treat the underlying network, rather than the individual account, as the unit of accountability — a scammer who creates a replacement page after one is removed has not been meaningfully deterred, only inconvenienced, unless the infrastructure itself, per Chapter 16's event-history architecture, makes the replacement traceable to the same accumulated record as the original. And users should retain meaningful control over how much synthetic interaction they wish to engage with, consistent with Chapter 9's argument that the goal is not homogenizing the environment into a single verified-human standard, but making the choice Chapter 9 described — currently made silently, on the user's behalf, by systems under no obligation to disclose that a choice has been made at all — into one the user makes knowingly.

17.4. What This Requires of Measurement Systems Internally

None of Section 17.3's commitments survive as anything more than a public statement unless they are embedded in the internal measurement systems, promotion criteria, and resource allocation decisions that Chapter 5 showed actually govern day-to-day institutional behavior — not the mission statement, but the dashboard a product manager is evaluated against. This is where the Correction Suppression Principle from Chapter 4 makes its final, most direct appearance in this book: the costs Chapter 7 documented as currently externalized onto individual users must be recognized, internally, as costs attributable to specific design choices, with accountability for those costs located at the institutional level where the relevant choices are actually made — not distributed, per Chapter 6, across a lateral network of teams none of which was assigned responsibility for the systemic condition their combined local decisions produced.

17.5. The Recursion Problem, One Last Time

This book cannot close, honestly, without returning to the claim Chapter 4 made and never retracted: no corrective system stands outside the geometry of correction, including whatever institutions eventually take up the proposals in this chapter. A regulator empowered to enforce the legibility standards of Section 17.3 is itself a system with its own evaluation function, and Chapter 4's argument applies to it with exactly the same force it applied to the platforms this book has spent sixteen chapters examining. A regulator can come to measure its own success by enforcement actions filed rather than by harm actually reduced. A stan-

dards body can come to measure its own success by the existence of a published standard rather than by whether platforms' actual behavior changed in response to it.

This book does not propose an exemption from this dynamic, because Chapter 4 already established that none is available. What it proposes instead, consistent with Chapter 14's argument for sympoiesis over centralization, is that whatever institutional apparatus eventually enforces the commitments in this chapter should itself be built on the same principles this book has argued for throughout: distributed rather than singular authority, so that no one point of capture is sufficient to compromise the entire mechanism; persistent, auditable histories of its own enforcement decisions, so that its own record remains as legible and contestable as the one it demands of the platforms it oversees; and structural openness to challenge from the human participants whose lives its decisions actually shape, rather than a closed, self-referential evaluation of its own success. A regulator that exempted itself from these requirements would not be a correction. It would be, in the precise sense this book has used the term throughout, the pathology recurring one level up — and the only honest response to that risk is not a promise that it cannot happen, but an architecture, per Chapter 16, in which it remains visible and contestable if it does.

17.6. Instantiating the Regulator's Own ρ

The preceding section states what a non-exempt corrective institution requires. Stated only that way, however, it risks being exactly the kind of claim Chapter 16 refused to make about platforms without a mechanism attached — that a system "should" anchor to something external, without specifying the anchor precisely enough that a reader could check whether a given institution actually satisfies it. Chapter 16 did not stop at saying platforms should preserve identity provenance and human-anchored meaning; it specified $\mathcal{E}$, $I(a)$, and the quadrangulation inequality concretely enough to be checked against. This section owes the regulator the same treatment.

A regulatory or standards body's own enforcement decisions can themselves be written as an event history, $\mathcal{E}_{\text{reg}} = \{e_1, e_2, \ldots\}$, structurally identical to the Sphere-pop architecture of Chapter 16: each finding, enforcement action, or certification the body issues is an irreversible event, and the body's own accumulated record, $I(\text{reg})$, cannot be quietly edited or reset any more than a platform's user history can under that architecture. A regulator that reverses a prior finding does not

erase the earlier one; it adds a new event that stands alongside it, permanently reviewable by anyone auditing the body's history rather than only by the body itself.

This addresses identity provenance for the regulator, in Chapter 16's sense. It does not yet address the second, independent channel Chapter 16 insisted was not reducible to the first: whether the regulator's own meaning-stabilizations — "this platform is compliant," "this design change is sufficient" — remain quadrangulated rather than closing among the regulator's own staff and appointees. Applying Chapter 16's constraint literally: a stabilization event e_{stab} asserting compliance is valid only if $\Phi(\text{reg}, O) \approx \Phi(h, O)$ for some human perspective h within the required tolerance ϵ. The entire force of this constraint depends on how h is specified, and this is the point at which most real institutions quietly fail without ever violating the letter of an oversight requirement: if h is drawn from the regulator's own staff, its political appointees, or a panel it selects and can replace, the constraint is satisfiable by construction, and ρ with respect to h approaches a maximum trivially, because the regulator has simply arranged to agree with itself. A corrective mechanism whose own anchor is self-selected has not raised ρ with respect to any channel it does not control; it has restated the self-referential condition from Chapter 3 one level up, dressed as compliance.

The requirement this book proposes, in place of a self-selected h, is that h be drawn from a rotating, independently constituted body of the people the regulator's decisions actually affect — composition determined by criteria the regulator itself does not set unilaterally and cannot revise unilaterally once set, membership rotating on a fixed schedule the regulator cannot accelerate or delay, and empowered specifically to reopen a stabilized finding, per the quadrangulation constraint of Chapter 16, rather than merely to comment on it after the fact. This is not a claim that such a body guarantees ρ remains high indefinitely — Chapter 4 already ruled out any such guarantee, for any system, including this one. It is a claim about what distinguishes an anchor that could, in principle, actually falsify the regulator's self-assessment from an anchor that cannot, structurally, ever do so. The difference between those two conditions is the entire distance between the aspirational paragraph this chapter could have stopped at three paragraphs ago, and an architecture a skeptical reader could actually check an institution against.

17.7. What This Book Does Not Establish

Before closing, it is worth being direct about several things this book has not done, rather than let a reader discover the gaps unassisted or mistake the book's confidence of argument for a claim of completeness it has not earned.

This book has not established that waymaking, in the sense Chapter 12 developed, is measurably eroded by platform design in a way any existing study has demonstrated. The mechanism is derived from first principles — friction sustains negotiation, negotiation sustains reality-contact — and is offered as plausible and consistent with the rest of the framework, not as an empirically settled finding. A skeptical reader is right to want controlled evidence this book does not supply, and right to notice that "waymaking," defined broadly enough to span a single moment of attention and a lifetime's intellectual project, is difficult to falsify with any single study design. That difficulty is a real limitation of the claim as stated, not merely of the evidence gathered for it.

This book has not established, with real institutional case studies, that sympoietic correction actually degrades more gracefully than centralized correction under sustained adversarial pressure. Chapter 14's argument is structural and follows from Chapter 4's recursion result; it is not backed here by a comparative history of, for instance, how open-source governance or Wikipedia's edit-dispute mechanisms have actually fared over decades against exactly the capture dynamics this book predicts they should resist better than a centralized alternative would. Such histories exist, are informative, and are not engaged with here.

This book has not established a detailed neurobiological account of why subcortical prediction error, in Panksepp's sense, is structurally incapable of resolution through the ordinary belief-updating Chapter 10 describes for cortical error. Chapter 11's treatment of jouissance is offered as a suggestive redescription of a real phenomenon in the vocabulary this book has built, not as a mechanism specified in enough biological detail to generate testable predictions of its own.

And, as Chapter 4 already conceded directly, this book has not supplied an estimator for ρ — a way of turning the diagnostic orientation this framework offers into a number a real auditor could compute and defend.

None of these are small gaps, and treating them as items for a future research program rather than pretending to have closed them here is a deliberate choice, not an oversight discovered only in retrospect. A book that tried to supply controlled evidence for waymaking, comparative institutional histories for sympoiesis, a

complete neurobiological mechanism for jouissance, and an operational estimator for ρ would be several books, each requiring expertise and evidence this one does not marshal. What this book offers instead is the framework within which those separate research programs could be pursued, and pursued by people better positioned than a single author working across four registers to do any one of them properly. That is a real limitation. It is also, this book would argue, a more honest place to stop than false completeness would have been.

17.8. The Redesign of Trust

The experience this book opened with — the impersonation account, the vanished pet deposit, the low-grade suspicion that has become, for most people who use large platforms, simply part of what using them feels like — is not, on the account this book has developed, a failure of technology to keep pace with malicious actors. It is the emergent property of systems whose evaluation, at every scale from a single recommendation algorithm to the corporate governance structures that oversee it, has drifted toward self-reference under ordinary optimization pressure, without any single decision being made to abandon the correction that would have prevented it.

Rebuilding trust in these environments does not require eliminating synthetic activity, which this book has argued throughout is neither possible nor, in its consensual forms, desirable. Nor does it require a retreat from networked sociality into some imagined, fully authentic communicative past that likely never existed in the unambiguous form nostalgia for it usually assumes. It requires, instead, the construction of informational legibility as a public good — systems in which the provenance, nature, and coordination of communicative signals remain visible to the people who encounter them, and in which the systems built to guarantee that visibility remain, themselves, visible and contestable in turn. That is the difference between an environment a person can navigate with accurate information about what they are actually encountering, and one they must inhabit under a suspicion they have learned to carry because the information that would let them set it down has been, whether by design or by the ordinary drift this book has spent its length describing, withheld. Closing that distance is not a technical problem, finally, or even only an institutional one. It is the same problem, recurring at every scale this book has examined, and it has, at every scale, the same shape: not the absence of a channel back to reality, but the discipline, never permanently secured, of remaining willing to consult it.

A Note on What Follows

The appendices below supply the formal machinery referenced, but not fully derived, in the body chapters. Each is tied explicitly to the chapter it supports, and none introduces content the corresponding chapter's argument depends on beyond what that chapter already states — the appendices exist to let a technically inclined reader see the derivations in full, not to smuggle in additional claims the main argument requires but never made explicit.

One appendix, Appendix F, carries a different status than the others and deserves a word of explanation. The Relativistic Scalar–Vector Plenum (RSVP) formalism appears in the body of this book only in Chapter 12's discussion of waymaking, where it supplies a single equation of motion for an agent's trajectory through an informational field. It is treated here as an appendix rather than expanded into a full chapter, on the following reasoning: the book's central argument, developed across Chapters 4 through 17, does not require RSVP's specific field-theoretic machinery, and every claim that appendix supports could, in principle, be restated without it. It is retained because it gives the waymaking formalism a physically motivated grounding some readers will find illuminating, and because it connects this book's argument to a broader research program the author has developed elsewhere. A reader uninterested in that connection can skip Appendix F entirely without losing anything Chapters 1 through 17 depend on.

Appendix A

Platform Engagement Optimization

Supports Chapter 3.

Let $U = \{u_1, u_2, \ldots, u_n\}$ denote the set of users and $C = \{c_1, c_2, \ldots, c_m\}$ the set of content artifacts. The platform recommendation system is a mapping $R : U \times C \to [0, 1]$, where $R(u, c)$ is the predicted engagement probability for user u and content c. The platform's objective, in practice, is

$$\theta^* = \arg\max_{\theta} \sum_{u,c} R_\theta(u, c)$$

subject to operational constraints. Synthetic agents become advantageous to this objective precisely when their contribution to engagement approaches that of genuine human activity, $A_{\text{synthetic}} \approx A_{\text{human}}$, producing the equilibrium condition referenced in Chapter 3:

$$\frac{\partial A}{\partial \text{authenticity}} \approx 0.$$

This is the formal correlate, at the level of a single platform's optimization, of the general claim Chapter 4 states for any evaluative system: when a channel does not appear in the objective being maximized, the system's behavior becomes insensitive to that channel regardless of the sophistication with which the objective is otherwise pursued.

Appendix B

The PRMO Framework

Supports Chapter 15.

The PRMO framework decomposes subjectivity into Subj $= (P, R, M, O)$: perception, representation, meaning, and the real — a label deliberately distinct from Chapter 4's $S(x)$, which denotes a system's internally generated signals and is not the same kind of object as this decomposition. Each artificial system realizes a projection π : Subj $\rightarrow$ Subj$_i$ onto some subset of these dimensions. Current large language models realize Subj$_{\text{LLM}} = (R)$ alone. Embodied AI systems realize Subj$_{\text{AS}} = (P, R)$. Synthetic social systems, of the kind Part I of this book examined, realize Subj$_{\text{SyS}} = (P, R, M)$. The real supplies a grounding constraint, O : Subj $\rightarrow \mathbb{R}^n$, that is not itself reducible to any of the other three dimensions — and which is the same O, in role if not in formal derivation, that Chapter 4 leaves generic and Chapter 10 instantiates psychoanalytically.

Ordinary meaning formation emerges through triangulation: $M = f(P_1, P_2, O)$, the convergence of two perspectives against a shared grounding channel. Quadrangulation, developed in full in Chapter 16, extends this to $M = f(P_h, P_{a1}, P_{a2}, O)$, where P_h denotes a human participant whose perspective cannot be eliminated from the meaning-formation process without invalidating the resulting field's coherence. Formally, if

$$\|\Phi(a_1, O) - \Phi(h, O)\| > \epsilon,$$

the interpretive process must reopen rather than proceed to a stabilized conclusion — the constraint referenced directly in Chapter 16's Section 16.3.

Appendix C

Predictive Coding and Variational Free Energy

Supports Chapters 10 and 11.

Under the Free Energy Principle, an agent maintains an approximate posterior belief $q(z)$ over hidden states z given observations s. Variational free energy is defined as

$$F = D_{\mathrm{KL}}\big(q(z) \,\|\, p(z|s)\big) - \ln p(s).$$

Prediction error, $\epsilon = s - \hat{s}$, is weighted by a precision term π that determines how strongly a given error propagates through the predictive hierarchy and compels revision. Surplus prediction error — error that persists despite repeated updating, discussed in Chapter 11 as the computational correlate of Lacanian jouissance — can be written

$$J \sim \pi\, \epsilon_{\text{persistent}},$$

occurring specifically when the generative model is structurally incapable of reducing a given error through the ordinary mechanism of belief updating. An engagement-optimized platform, per Appendix A's formalism, amplifies exactly the high-precision error signals this equation describes, systematically, across its user population — instrumentalizing a cognitive mechanism at scale that evolved for entirely different, and generally adaptive, purposes.

Appendix D

Event Histories and Spherepop Structures

Supports Chapter 16 (Section 16.2).

Spherepop treats interaction as an irreversible event history $\mathcal{E} = \{e_1, e_2, \ldots, e_t\}$ (a script letter, kept distinct from the evaluation function E of Chapter 4), with each event $e_i = (a_i, t_i, \sigma_i)$ recording the acting agent a_i, timestamp t_i, and state transformation σ_i. State evolves recursively:

$$\Sigma_{t+1} = \Sigma_t + \sigma(e_t), \qquad t_i < t_j \implies e_i \prec e_j.$$

Agent identity is constituted by the accumulated record $I(a) = \{e_i : a_i = a\}$ rather than by an externally assigned, resettable profile score. The four core operators — Pop, Refuse, Bind, and Collapse — each produce a new event within $\mathcal{E}$, preserving irreversibility by construction rather than by policy. In the context of Appendix B's formalism, a Spherepop event structure implements the real (O) directly: a causally anchored substrate grounding identity and meaning in durable history rather than in manipulable symbolic representation.

Appendix E

Quadrangulation as a Constraint System

Supports Chapter 16 (*Section 16.3*).

Let a sense-field consist of agents $A = \{h, a_1, a_2, \ldots\}$, with h a human subject. Meaning stabilization occurs through $\Phi : A \times O \to \mathbb{R}$. Triangulation stabilizes meaning when $\Phi(a_1, O) \approx \Phi(a_2, O)$. Quadrangulation imposes the additional constraint

$$\Phi(a_1, O) \approx \Phi(a_2, O) \approx \Phi(h, O).$$

If $\|\Phi(h, O) - \Phi(a_1, O)\| \geq \epsilon$, the system must reopen the interpretive process rather than treat the machine-agent consensus as valid. Mapped onto the Spherepop formalism of Appendix D: a meaning-stabilization event e_{stab} is valid only if it is consistent with a non-trivial projection of $I(h)$. Machine-only coherence, lacking this anchor, is formally incomplete — not merely unverified, but excluded by construction from counting as a valid stabilization at all.

Appendix F

The Relativistic Scalar–Vector Plenum

Supports Chapter 12. See the note at the head of this section regarding this appendix's status relative to the others.

The RSVP framework models informational structure as three coupled fields: scalar density $\Phi(x,t)$, vector flow $\mathbf{v}(x,t)$, and entropy $\mathcal{S}(x,t)$, forming the state vector $\Psi = (\Phi, \mathbf{v}, \mathcal{S})$.

Field Equations

$$\frac{\partial \Phi}{\partial t} + \nabla \cdot (\Phi\, \mathbf{v}) = D_\Phi \nabla^2 \Phi - \alpha \mathcal{S},$$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla \Phi + \nu \nabla^2 \mathbf{v} - \beta \nabla \mathcal{S},$$

$$\frac{\partial \mathcal{S}}{\partial t} + \mathbf{v} \cdot \nabla \mathcal{S} = D_\mathcal{S} \nabla^2 \mathcal{S} + \gamma |\nabla \Phi|^2.$$

Relation to Waymaking (Chapter 12)

An agent's trajectory through this field is written $\gamma(t) \subset \Psi$, with dynamics

$$\frac{d\gamma}{dt} = -\nabla \mathcal{S} + \eta,$$

where η represents the agent's own exploratory variation, referenced directly in Chapter 12's Section 12.3 and Chapter 13's Section 13.3. When an external guiding function $A(\gamma, t)$ — the platform's optimization layer, discussed in Chapter 13 — comes to dominate the dynamics, exploratory variance is suppressed toward zero and the agent's trajectory is replaced by one externally optimized for engagement rather than for the agent's own understanding.

Quadrangulation in the Plenum

In a multi-agent environment, each agent observes a projection $\Psi_i = \Pi_i(\Psi)$. Ordinary alignment requires $\|\Psi_i - \Psi_j\| < \epsilon$. Quadrangulation, per Appendix E, imposes the additional constraint $\|\Psi_i - \Psi_h\| < \epsilon$, where h denotes a human observer. Machine-only coherence corresponds to a field configuration that has drifted beyond the accessible region of $\Pi_h(\Psi)$ — stable within the plenum, in the sense of satisfying the field equations above, but disconnected from the human observer's own generative model of the same field.

Bibliography

[1] Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348, no. 6239 (2015): 1130–1132.

[2] Boyd, danah. *It's Complicated: The Social Lives of Networked Teens*. New Haven: Yale University Press, 2014.

[3] Dall'Aglio, John. *A Lacanian Neuropsychoanalysis: Consciousness Enjoying Uncertainty*. The Palgrave Lacan Series. Cham: Palgrave Macmillan, 2024. doi:10.1007/978-3-031-68831-7.

[4] Debord, Guy. *La Société du spectacle*. Paris: Éditions Buchet-Chastel, 1967. English translation: *The Society of the Spectacle*. New York: Zone Books, 1994.

[5] Freud, Sigmund. *Beyond the Pleasure Principle*. London: Hogarth Press, 1920.

[6] Friston, Karl. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11 (2010): 127–138.

[7] Friston, Karl. "Life as We Know It." *Journal of the Royal Society Interface* 10, no. 86 (2013): 20130475.

[8] Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press, 2018.

[9] Gloy, Johann Flemming, and Simon Olsson. "HollowFlow: Efficient Sample Likelihood Evaluation Using Hollow Message Passing." In *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. `https://openreview.net/forum?id=KYlIC6sLhw`.

[10] Jelinek, Thorsten, Patrick Glauner, Alvin Wang Graylin, and Yubao Qiu. "The Post-Turing Condition: Conceptualising Artificial Subjectivity and Synthetic Sociality." 2026. `https://arxiv.org/abs/2601.12938`.

[11] Lacan, Jacques. *Écrits: The First Complete Edition in English*. Translated by Bruce Fink. New York: W.W. Norton, 2006.

[12] Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

[13] Panksepp, Jaak. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford: Oxford University Press, 1998.

[14] Pariser, Eli. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press, 2011.

[15] Postman, Neil. *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. New York: Viking Penguin, 1985.

[16] Solms, Mark. *The Hidden Spring: A Journey to the Source of Consciousness*. New York: W.W. Norton, 2021.

[17] Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review* 5, no. 3 (1997): 305–321.

[18] Tufekci, Zeynep. "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency." *Colorado Technology Law Journal* 13, no. 2 (2015): 203–218.

[19] Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The Spread of True and False News Online." *Science* 359, no. 6380 (2018): 1146–1151.

[20] Wu, Tim. *The Attention Merchants: The Epic Scramble to Get Inside Our Heads*. New York: Knopf, 2016.

[21] Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.