Intelligence as Perspectival Convergence

Flyxion

Contents

A Note on Structure

This book intentionally presents several frameworks that are later revised, corrected, or abandoned. The first model of elimination (Chapter 3) turns out to be unrepairable and is replaced in Chapter 4. The first account of how a child's vocabulary develops (Chapter 7) turns out to be wrong about what a "subsystem" is, and is repaired mid-chapter. A clean curriculum hypothesis in Chapter 11, and a clean independence hypothesis in Chapter 12, both meet the same fate. This is not because the manuscript was undecided about its own claims while being written, and a reader encountering the pattern cold could be forgiven for suspecting otherwise.

The central claim of this book is that concepts become trustworthy not by being declared correct, but through repeated cycles of proposal, criticism, elimination, and repair — and that the difference between a hard core and a soft periphery only becomes visible once something has been tested against reality and found wanting. A book that only ever showed its readers the final, successful version of each idea would be describing that process from the outside. This one tries, instead, to put the reader inside it: each provisional model is built in earnest, allowed to fail on its own terms, and then repaired in a way the reader can check against the failure that prompted it.

To make this easier to track while reading, provisional models that are later revised are marked in a shaded box like this one is not, labeled PROVISIONAL MODEL, and the repair that replaces each one is marked in a second, differently shaded box labeled REPAIR. A box of either kind signals exactly what it is: a stage in an argument, not the book's last word on the subject. The book's actual closing position on any given question is always the content of the final repair box in the relevant chapter, or, failing that, the surrounding prose once the boxes stop appearing.

One further piece of apparatus is marked differently still. Section 17.4 builds a small amount of physics-flavored formal machinery — a Lagrangian, an Euler–Lagrange equation, a Hamiltonian — to describe, speculatively, how the book's elimination dynamics might behave as a continuous process. This material is explicitly conjecture-level, clearly subordinate to the operational definitions that precede it in the same section, and is marked as an OPTIONAL FORMAL EXTENSION: a reader uninterested in the dynamical apparatus can skip the boxed material entirely without losing anything the rest of the book depends on.

Why this structure was chosen, rather than the more conventional alternative of presenting only the finished framework, is explained directly in the book's final chapter, once the

reader has the full pattern in view to judge for themselves whether it worked.

PART I

The Problem of the Definition

1

The Trouble with Defining Intelligence by Discovery

In plain terms: Every generation confidently redefines intelligence, and every next generation finds the definition wrong. That cycling, rather than converging, is itself evidence that intelligence isn't a single hidden thing waiting to be correctly named. This chapter states that puzzle precisely; it doesn't yet solve it.

Ask a roomful of researchers to define intelligence and you will get a roomful of different answers, and this would not be troubling if the answers were converging — if each decade's definition were a refinement of the last, narrowing in on some fixed target the way successive measurements of a physical constant narrow in on its true value. They are not converging. They are cycling.

Intelligence has been defined, at various points and by serious people, as the capacity to optimize a reward signal, as the capacity to predict future sensory states, as the capacity to compress experience into efficient representations, as a form of computation, as the manipulation of symbols according to rules, as an emergent property of sufficiently complex information processing, as adaptive behavior in a changing environment, and as whatever a particular test happens to measure. Each definition arrived with the confidence of a discovery — *this* is what intelligence actually is, beneath the folk confusion — and each was eventually absorbed, qualified, or quietly abandoned by the next generation, which then proposed its own discovery with the same confidence.

This pattern deserves a name, because once named it becomes hard to unsee: call it the *discovery cycle*. A theory is proposed as if intelligence were a fixed target being progressively triangulated. The theory illuminates real cases — usually the cases its proponents had in mind when constructing it — and fails on cases nobody had in mind. The failures accumulate. A new theory is proposed, again as a discovery, again illuminating a different set of cases and failing on others. Nothing about this process looks like convergence on a stable target. It looks like something else.

The most natural explanation — and the one this book will spend its length defending, complicating, and ultimately earning the right to state precisely — is that "intelligence" was

never a fixed target in the first place. Not because there is nothing real underneath the word, but because what is underneath the word is not the kind of thing a single definition is built to capture: not a latent variable waiting to be measured more precisely, but a structured, multidimensional pattern that different observers, working from different vantage points and against different failure cases, partially and unevenly perceive. Each definition is not wrong so much as a true description of one face of something that has several.

If that is right, then the discovery cycle is not a series of failures to find the right definition. It is the visible signature of a different process entirely — one in which a concept gets *built*, not found, through repeated contact between candidate descriptions and the world's resistance to them. This book calls that process *perspectival convergence*, and the rest of Part I exists to state the problem precisely enough that the machinery built in Part II has something exact to do.

Notation

Let

$$\Theta = \{\theta_1, \theta_2, ...\}$$

denote the space of candidate definitions of intelligence — every precise proposal that has been or could be advanced, from "intelligence is the capacity to maximize expected reward" to "intelligence is whatever an IQ test measures" to far more recent and far stranger candidates. Equip Θ with a conceptual-distance metric $d_\Theta(\theta_i, \theta_j)$, a rough and for now informal measure of how far apart two candidate definitions are in the space of things they would classify as intelligent or not.

The discovery cycle, restated in this notation, is the observation that the historically realized sequence of candidate definitions does not look like a sequence converging toward a single θ^*. It looks, across the twentieth and early twenty-first centuries, like $|\Theta_t|$ — the number of seriously defended candidates in active circulation — has tended to grow rather than shrink, even as individual theories have died.

Conjecture 1.1 (Definition-Space Expansion)**.** In the historical record of intelligence research, $|\Theta_t|$ increases for long stretches before any convergence mechanism causes it to decrease.

This is stated as a conjecture rather than an established fact because it has not, in this book, been checked against an actual census of the literature — it is offered as the intuition that motivates everything that follows, and a reader who suspects it is false (perhaps definitions really have been narrowing, just slowly, and the appearance of expansion is an artifact of which decades get remembered) should treat the rest of the book as conditional on it being roughly right. What the conjecture is for is this: if true, it means the field needs an elimination mechanism — some way of taking the expanding set Θ_t and producing a narrower, more defensible Θ_T — and it means that mechanism cannot simply be "wait for more

research," since more research, on the historical evidence, has been expanding Θ rather than shrinking it. Chapter 3 builds the first candidate for such a mechanism. It will turn out, in Chapter 4, to be the wrong one.

2

Two Ways a Concept Can Be Fixed

In plain terms: Is intelligence something real we're discovering, or something we're collectively inventing through argument? This chapter stakes out a middle position: part of what "intelligence" means is forced on us by reality and won't budge; part is just convention and could have gone differently. The tools to tell which part is which come later.

Before building any machinery, it is worth being precise about what success would even look like — what it would mean for the elimination process to work, and what kind of thing the survivor would be once it had worked. There are two very different answers available, and the difference between them is not a technicality. It determines whether this entire project is a contribution to epistemology (we are getting closer to a truth that was already there) or to something closer to sociology of concepts (we are watching a category get built, with no further fact about what it "really" is once the building is done).

Discovery. On this view, there is a real, mind-independent fact about what intelligence is — a θ^* somewhere in Θ — and the elimination process, whatever form it takes, is a method for approximating θ^* more and more closely. Mistakes are possible (a definition can be wrongly eliminated, or a bad one wrongly retained), and the measure of the process's success is how close the eventually-converged Θ_T comes to containing or approximating θ^*.

Constitution. On this view, there is no θ^* waiting to be found. The word "intelligence" does not refer to a pre-existing natural kind; it refers to whatever survives the elimination process, full stop. Θ_T does not approximate anything external to itself — it *is* the thing the word picks out, because the elimination process is not measurement, it is concept-formation. There is no further question to ask, on this view, about whether Θ_T got it right, because "getting it right" presupposes exactly the fixed external target that constitution denies exists.

These positions are not merely different emphases. They make different predictions about what should happen if the elimination process were rerun from a different starting point, by a different culture, with a different sequence of perspectives encountered. Discovery predicts convergence to (approximately) the same Θ_T regardless of path, because there is a real target pulling every honest inquiry toward it. Constitution predicts no such guarantee — a differently-ordered, differently-populated elimination process could land somewhere else entirely, with no sense in which one outcome is more correct than the other.

The working position

This book does not adopt either pole. It adopts what will be called *weak realism*, and stating it precisely now — even before the machinery exists to justify it — sets the target for everything that follows.

Definition 2.1 (Hard/Soft Decomposition)**.** Let

$$\Theta_t = H_t \cup S_t$$

where H_t (the *hard core*) is the portion of the current candidate set fixed by content that has survived contact with resistant, non-negotiable reality — a system's measurable incapacity, a hazard that reliably produces the outcome it threatens, a counterexample that is simply, checkably true — and S_t (the *soft periphery*) is the portion fixed only by local, revisable, social agreement: emphasis, caricature, rhetorical amplification, cultural priority.

Conjecture 2.2 (Differential Stability)**.** As evidence accumulates, $dH_t/dt \to 0$ — the hard core stabilizes and stops moving, because what fixes it (contact with reality that doesn't change its mind) doesn't go away. dS_t/dt, by contrast, may remain nonzero indefinitely — the soft periphery can keep shifting forever, because nothing forces it to stop.

This is stated as a conjecture rather than a theorem because nothing in this chapter yet defines what "evidence accumulation" formally does to H and S — that requires the elimination machinery of Chapter 5, in particular the fidelity functional that will let the book distinguish, for any given candidate constraint, whether it belongs to the kind of evidence that stabilizes H or the kind that merely reshuffles S. Until that machinery exists, the hard/soft split is an organizing intuition, not a derived result. It is the intuition the rest of the book is built to earn.

If the conjecture holds, weak realism follows naturally: intelligence is neither purely discovered (there is no single fixed θ^*, only a hard core that itself may be multidimensional rather than a single point) nor purely constituted (the hard core is not up for grabs — no amount of cultural rearrangement will make a system that provably cannot do X count as having done X). The interesting work, on this view, is not deciding the fork once and for all. It is learning, case by case, which parts of any given definition belong to H and which belong to S — which is exactly the question the elimination machinery of Part II is designed to answer, and exactly the question that, by the book's own logic, the book's own argument about intelligence must itself survive being asked of.

PART II

The Formal Machinery

3

The Space of Candidate Definitions

In plain terms: This chapter builds the simplest possible filter for candidate definitions: collect evidence from different sources and keep only what's consistent with all of it. It's a natural first move, marked here as provisional, since the next chapter proves it cannot work. Along the way, it also gives a precise name to the sources of evidence: a *perspective*.

The first attempt at an elimination mechanism is the most natural one available, and it is worth building carefully, in full, before the next chapter shows why it fails — because the way it fails turns out to determine the shape of everything that replaces it.

Before any elimination mechanism can be built, the thing doing the eliminating needs a name. This book is titled around perspectival convergence, and everything that follows talks constantly about observers, cultures, institutions, developmental stages, and rival theoretical schools as sources of evidence — yet none of these have so far been given a single formal home. That gap is closed now, briefly, so that the object the title refers to actually exists in the notation rather than remaining a suggestive word floating above it.

Definition 3.1 (Perspective)**.** A perspective P is any source capable of producing evidence that bears on which candidate definitions in Θ remain defensible — an individual observer, an interview subject, a culture's folk theory, an institution's recorded practice, a developmental stage, a scientific research program, or a system's demonstrated capacity or failure. A perspective need not be a person; it need only be a source from which a constraint can, in principle, be extracted.

Definition 3.2 (Induced Constraint)**.** Each perspective P_i induces a constraint on the candidate space via a map

$$\Phi : \{\text{perspectives}\} \to \{\text{constraints on } \Theta\}, \qquad C_i = \Phi(P_i).$$

Φ is not assumed to be simple, uniform, or fully specified in advance — extracting C_i from a given P_i is itself substantive interpretive work, and different chapters of this book amount to case studies in what Φ looks like for different kinds of perspective (an ethnographic record, an institutional history, a developmental dataset). What matters for the formalism is only that the output of Φ, whatever the input, is a constraint of the kind the rest

of this chapter and the next can operate on. With this in place, perspectives are no longer a word the book uses informally around the edges of its math — they are the explicit input to the function that produces everything Θ_n, μ_t, and later chapters' case studies are built from.

Criteria for perspectivehood

The definition above is broad enough to invite a fair question: is a microscope a perspective? An LLM? A culture? A three-year-old child? Left unanswered, the breadth of the definition risks becoming a license to call anything a perspective whenever doing so is convenient, which would make the entire elimination apparatus that follows unfalsifiable in the same way an uncontrolled subsystem split would be (a concern this book takes seriously enough to build an explicit anti-vacuity principle for, later, in Chapter 7). Four criteria, taken jointly, are offered to draw the line.

A perspective must be capable of *generating constraints* — it must produce, in principle, some output that bears on which candidates in Θ remain defensible, not merely exist as an object that could in theory be studied. A bare microscope generates no constraint by itself; a microscope used by a researcher to test a specific structural hypothesis, as part of Chapter 17's distinctive-features case, does. It must be *distinguishable from other perspectives* — there must be a principled way to say where one perspective ends and another begins, which is what makes pairwise independence (Chapter 17) a meaningful quantity to compute rather than an arbitrary partition imposed after the fact. It must be *capable of disagreement* — a perspective that is constitutionally incapable of producing a constraint inconsistent with any candidate contributes nothing to the elimination dynamics and is, for the purposes of this book, inert; this rules out perspectives defined so narrowly that they can only ever confirm what is already assumed. And it must be *independently describable* — nameable and specifiable without reference to the candidate it happens to be evaluated against, so that $\varepsilon_i(\theta)$ and $\mathrm{Ind}(i, j)$ in Chapter 17 can be assessed without circularity.

By these criteria, a culture, an institution, a developmental stage, and a research program all qualify, provided each can be shown (not merely asserted) to satisfy all four — which is exactly the burden Chapters 6 through 13 attempt to discharge case by case, rather than assuming met by default. An LLM qualifies under the same test only to the extent that its outputs can be shown to constrain Θ independently of the training data and assumptions a human perspective already supplied — a question this book does not resolve and flags explicitly as open, since the independence criterion is precisely what would need to be established, not assumed, before treating a model's output as a perspective in this technical sense rather than a restatement of perspectives already counted.

With Perspective and Induced Constraint in hand, the naive elimination mechanism can now be stated precisely.

Definition 3.3 (Naive Elimination)**.** Let n perspectives $P_1, \ldots, P_n$ be consulted in sequence, each inducing a constraint $C_i = \Phi(P_i) \subseteq \Theta$ — the subset of candidate definitions consistent with what that perspective has shown. Define

$$\Theta_n = \bigcap_{i=1}^{n} C_i.$$

This is elimination in its most literal form. Each new perspective rules out whatever it is inconsistent with, and the survivors are whatever remains consistent with everything encountered so far. It has an obvious appeal: it requires no weighting, no judgment calls about which perspectives matter more than others, no machinery beyond set intersection. Every constraint counts equally; every constraint that rules something out, rules it out permanently.

Proposition 3.4 (Monotonicity)**.** *For pure intersection,* $\Theta_{n+1} \subseteq \Theta_n$ *for every* n*.*

This follows immediately from the definition of intersection — adding one more set to an intersection can only remove elements, never add them — and the proof is not worth more than the sentence just given. It is stated explicitly because it is about to become the chapter's entire problem, and a problem is easiest to see once its source has been isolated and named. *This model is proven unworkable in Chapter 4 and repaired in Chapter 5.*

4

Non-Monotonic Revision

In plain terms: The previous chapter's model has a provable fatal flaw: if you only ever narrow your options and never let anything back in, a single bad piece of evidence permanently destroys your chances of reaching the right answer, no matter how much good evidence follows. This chapter proves that, then begins building a model that can recover from its own mistakes.

Monotonicity sounds, on first encounter, like a virtue — a guarantee that the elimination process never goes backward, never re-admits something already ruled out, always tightens. It is in fact the naive model's fatal flaw, and the flaw is provable rather than merely suspected.

Theorem 4.1 (Irreversible Elimination)**.** *If a sequence of updates is strictly monotonic and shrinking — $\Theta_{t+1} \subseteq \Theta_t$ for all t — then mistaken elimination of the true candidate θ^* can never be repaired by any subsequent update in the sequence.*

Proof. Suppose $\theta^* \in \Theta_t$ for some t, but a mistaken constraint causes $\theta^* \notin \Theta_{t+1}$. By monotonicity, $\Theta_{t+k} \subseteq \Theta_{t+1}$ for every $k \geq 0$. Since $\theta^* \notin \Theta_{t+1}$, and every later set is a subset of Θ_{t+1}, it follows that $\theta^* \notin \Theta_\tau$ for every $\tau \geq t+1$. No future state of the sequence contains θ^*. If the process is meant to eventually converge to a state containing the true candidate — which is the entire point of running it — this is a direct contradiction. The process cannot recover from a single mistaken elimination, ever, no matter how much further evidence accumulates.

This is, among the formal results in this book, the one that does the most work with the least apparatus. It does not depend on anything specific to intelligence, or to any particular domain — it is a structural fact about any process built on pure, monotonic intersection. And it has an immediate and uncomfortable consequence: every real elimination process is going to make mistakes. Constraints are sometimes wrong, sometimes based on incomplete information, sometimes the product of a biased or unrepresentative perspective. A model that can never recover from a single such mistake is not a model of how concepts should be revised. It is a model of how they get permanently corrupted by the first bad piece of evidence anyone happens to introduce.

The naive model must therefore be replaced — not patched, replaced — with something that allows re-entry: a process in which a previously down-weighted candidate can regain

plausibility if later evidence supports it. This requires moving from sets to something with graded membership.

Definition 4.2 (Belief Measure)**.** Let $\mu_t : \Theta \to [0, \infty)$ be a plausibility assignment over candidate definitions at time t, updated by some rule $\mu_{t+1} = U(\mu_t, C_t)$ in place of set intersection.

The naive model is the special case where U sends $\mu_t(\theta)$ to 0 whenever θ violates C_t, and leaves it unchanged otherwise — which is exactly what reproduces the monotonic, irreversible behavior just proven fatal. Any update rule that allows a previously-reduced $\mu_t(\theta)$ to subsequently rise again escapes Theorem 4.1, because the theorem's proof depends essentially on the impossibility of re-entry once excluded.
What such an update rule should actually look like — specifically, how much a given constraint should be allowed to move $\mu_t(\theta)$, and on what basis — is the question Chapter 5 exists to answer, where this repair is completed.

5

Fidelity: What Gives an Elimination Authority

In plain terms: Not all evidence deserves equal trust, and this chapter is the one the rest of the book depends on. It breaks "how much should I trust this piece of evidence" into four checkable questions and combines them into one number, *fidelity*. The central claim: people and cultures fail not by reasoning badly in general, but specifically by treating low-quality evidence as if it were high-quality.

A graded belief measure solves the irreversibility problem, but it immediately raises a harder one. If every constraint moves $\mu_t(\theta)$ by some amount, what determines that amount? Treating every perspective as equally authoritative is no more defensible than treating every perspective as infallible — a single offhand, biased, or rhetorically-constructed claim should not move belief by the same amount as a rigorously demonstrated counterexample. The naive model's flaw was irreversibility. Its less obvious but equally serious flaw, inherited by any careless graded replacement, is treating all eliminating evidence as if it carried the same weight. This chapter builds the machinery to stop doing that.

5.1 Four components of fidelity

Call the authority a given eliminating constraint deserves its *fidelity*, w_i. The claim of this chapter is that fidelity is not a single intuitive quantity but decomposes into four independently checkable components, each addressing a distinct way a piece of eliminating evidence can be untrustworthy even while looking, on its surface, exactly like a hard fact.

Causal directness (d_i). Does the eliminated possibility fail because of an intrinsic property of the thing itself, or because of an extended, contingent, multiply-mediated causal chain? A mathematical counterexample has maximal directness — the failure is definitional, with zero intervening contingent steps. A claim like "laziness leads to poverty" has low directness — it routes through labor markets, family circumstance, health, and luck, none of which the claim acknowledges.

Severity calibration (s_i). Does the outcome a constraint depicts match the outcome re-

ality actually produces, or is the depiction amplified beyond what the evidence supports? Formally, s_i is the ratio of actual conditional outcome probability to depicted outcome probability — close to 1 for a hazard that reliably produces what it threatens, far below 1 for a dramatized worst case presented as a near-certainty it is not.

Closure (c_i). Is the eliminated trajectory actually a closed door, or does the constraint silently omit real recovery paths that exist in the world but would weaken its rhetorical force if shown? High closure — meaning genuinely few or no real escape routes exist — should increase a constraint's fidelity, since the eliminated outcome really is as final as depicted. Low closure — many hidden recovery paths the constraint omits — should reduce fidelity, since the constraint overstates how final its eliminated outcome actually is.

Source independence (q_i). Would the elimination retain its force if asserted by a stranger with no social standing, or does its authority depend entirely on who is making the claim? A claim checkable by anyone, regardless of the speaker's status, has high source independence. A claim that only feels compelling because an elder, an institution, or a culture says so has low source independence.

Definition 5.1 (Fidelity Functional)**.** Combine the four components multiplicatively, each scaled to $[0, 1]$:

$$w_i = d_i \cdot s_i \cdot c_i \cdot q_i.$$

With this naming, every factor now points the same direction — larger is always better, larger always increases fidelity, and a reader need not mentally invert any single component to follow the formula. The multiplicative form is deliberate rather than incidental: it ensures that any single component collapsing toward zero drags the overall fidelity toward zero as well, rather than being averaged away by the other three. A constraint that is indirect, exaggerated, low-closure (full of hidden escape routes it doesn't show you), and dependent entirely on the speaker's authority should not score moderately on fidelity merely because it is, say, vivid. It should score close to zero, because every independent check it could pass, it fails.

Why the fidelity functional is multiplicative

This choice is worth justifying rather than leaving as a stylistic preference, because an additive functional was the more obvious first guess and was rejected for a specific reason. Consider the alternative,

$$w_i^{\text{additive}} = \tfrac{1}{4}(d_i + s_i + c_i + q_i).$$

Under this form, a constraint that is catastrophically authority-dependent ($q_i \approx 0$) but happens to score well on the other three components can still average out to a moderate, even respectable, fidelity score. That is exactly the wrong behavior for what fidelity is meant to represent. A constraint with zero source independence — one that is compelling only because of who is asserting it — should not be rescued by being otherwise well-calibrated and direct, because the entire question of whether it deserves authority *at all* turns on whether

it would survive being asserted by no one in particular. An additive form allows components to compensate for one another; a multiplicative form enforces a bottleneck, where the weakest link determines the overall strength of the chain rather than being smoothed over by an average. Since the four components were each motivated as addressing a distinct way a constraint can fail entirely independently of the others, a single catastrophic failure on any one of them should be allowed to be catastrophic for the whole, which is precisely what multiplication, and not addition, enforces.

5.2 The corrected update rule

With fidelity in hand, the belief measure can be updated by an amount proportional to how strongly a given constraint argues against a candidate, scaled by how much authority that argument deserves.

Definition 5.2 (Soft Elimination, corrected)**.** Let $e_t(\theta) \in [0, 1]$ measure how strongly the constraint encountered at time t argues against candidate θ. A first, naive version of the update would be

$$\tilde{\mu}_{t+1}(\theta) = \mu_t(\theta)(1 - w_t\, e_t(\theta)).$$

This expression is not yet adequate as stated, because repeated multiplicative shrinkage of this kind destroys probability mass without redistributing it — if μ is meant to remain a genuine probability distribution over Θ, summing to 1 at every step, this update will generally cause that sum to drift below 1 as t grows, since mass is removed at each step but never returned anywhere. The corrected version renormalizes:

$$Z_t = \sum_{\theta \in \Theta} \mu_t(\theta)(1 - w_t\, e_t(\theta)), \qquad \mu_{t+1}(\theta) = \frac{\mu_t(\theta)(1 - w_t\, e_t(\theta))}{Z_t}.$$

This is the update rule the rest of the book assumes whenever μ_t is invoked. It is worth pausing on why the correction matters beyond mathematical tidiness: every later argument in this book that reasons about μ_t in probabilistic terms — comparing likelihoods, talking about expected error, treating $\mu_t(\theta)$ as something like a degree of belief that should behave the way degrees of belief behave — silently depends on μ_t actually being a probability distribution at every step. The uncorrected version is not one. The corrected version is, and the next result confirms it stays one under repeated application.

Theorem 5.3 (Fidelity Stability)**.** *If $0 \leq w_t e_t(\theta) \leq 1$ for every θ and every t, and $Z_t > 0$ at every step, then $\mu_t(\theta) \geq 0$ and $\sum_\theta \mu_t(\theta) = 1$ for all t — that is, μ_t remains a valid probability distribution throughout the process.*

Proof. By induction on t. Base case: μ_0 is given as a valid probability distribution by construction. Inductive step: suppose μ_t is a valid probability distribution. Since $0 \leq w_t e_t(\theta) \leq 1$ for every θ, each factor $(1 - w_t e_t(\theta))$ lies in $[0, 1]$, so the product $\mu_t(\theta)(1 - w_t e_t(\theta))$ is non-negative

for every θ. Dividing by $Z_t > 0$ preserves non-negativity. For the sum:

$$\sum_\theta \mu_{t+1}(\theta) = \sum_\theta \frac{\mu_t(\theta)(1 - w_t e_t(\theta))}{Z_t} = \frac{1}{Z_t} \sum_\theta \mu_t(\theta)(1 - w_t e_t(\theta)) = \frac{Z_t}{Z_t} = 1.$$

So μ_{t+1} is non-negative and sums to 1: a valid probability distribution. By induction, this holds for all t.

This is a modest theorem — it confirms that a corrected, sensible-looking update rule behaves the way a sensible update rule should — but it is not vacuous. The original, uncorrected version of the same idea did not have this property, and a great deal of downstream reasoning in this book implicitly requires that it does.

5.3 Fidelity is not correctness

A misunderstanding worth heading off before the rest of this book builds further on this chapter: fidelity and correctness are not the same thing, and nothing in the apparatus above guarantees they coincide.

Fidelity measures how much authority a constraint deserves *if* it is taken seriously as evidence. It does not measure whether the constraint's content is actually true. A high-fidelity elimination can still be wrong: a carefully designed experiment can be confounded in a way no one yet detects; a mathematical proof can contain a subtle, undiscovered error; a witness with every mark of reliability can misremember a crucial detail. None of these possibilities are excluded by w_i being close to 1. Symmetrically, a low-fidelity elimination can accidentally be correct: a rumor can happen to predict a true outcome; a caricature can, despite its distortion, point at a genuine weakness; an authority-dependent claim asserted for bad reasons can still describe the world accurately. Low w_i does not entail falsehood any more than high w_i entails truth.

This distinction matters because a great deal of ordinary epistemic failure consists precisely in conflating the two. To treat a constraint's fidelity as though it settled the constraint's truth is to skip a step that this chapter's entire apparatus was built to keep visible. What the fidelity functional governs is a narrower and more tractable question: given that perfect, oracle-level knowledge of truth is not available to any real elimination process, how should authority be *apportioned* across constraints of varying reliability, so that belief revision tracks truth on average even though it cannot track it with certainty case by case? The goal of the machinery in this chapter is therefore not certainty but error management: high-fidelity constraints should, in the aggregate and over repeated use, produce fewer mistaken eliminations than low-fidelity ones. Whether they reliably do so for any particular candidate, on any particular occasion, remains an open and separate question — and is exactly the content of the Misweighted Elimination conjecture below.

5.4 What fidelity weighting is actually for

The deepest sentence in this book, the one that will recur in nearly every subsequent chapter under different names and in different domains, can now be stated precisely rather than informally: *intelligence-as-self-description fails when low-fidelity counterfoils are mistaken for hard eliminations.* A culture, an institution, or an individual reasoner that treats a rhetorically amplified, low-directness, low-closure, authority-dependent constraint as if it carried the weight of a mathematical counterexample is not reasoning poorly in some vague sense. It is making a specific, namable, structurally identical error every time, regardless of the domain in which it occurs — and this book's later chapters on institutions, on developmental pedagogy, on creativity, and on borrowed authority in ideology and moral panic are, formally, the same mistake examined in five different costumes.

What remains genuinely open, and is worth stating plainly rather than implying it has been settled: this chapter has built the machinery to *assign* fidelity once its four components are known, but it has not yet shown that doing so reliably reduces error in the resulting belief state — only that it is the principled way to try.

Conjecture 5.4 (Misweighted Elimination)**.** Let $L = \mathbb{E}[(\hat{\mu} - \mu^*)^2]$ denote the expected squared error between the model's belief state and the (unknown) true plausibility assignment. For constraints whose fidelity weights are correctly calibrated to their actual reliability,

$$\frac{\partial L}{\partial w} < 0$$

— increasing the authority given to a high-fidelity constraint reduces expected error, while doing the same for a low-fidelity constraint increases it.

This is precisely stated and, in principle, falsifiable — but it is not derived here. Deriving it would require an explicit model relating $e_t(\theta)$ to the unknown μ^*, which this book has not constructed. It is offered as the chapter's central open target rather than a result already in hand: the formal proof, should it exist, that calibrating elimination authority correctly is not merely intuitively appealing but provably error-reducing. Closing this gap is, of everything left open in this book, probably the single most valuable piece of further work available — and it is flagged here, rather than smuggled past as already accomplished, because a book whose subject is the difference between earned and borrowed authority cannot afford to borrow any of its own.

PART III

How Convergence Actually Behaves

6

Institutions: The Room for the River Case

In plain terms: Does a multi-part system (here, Dutch flood infrastructure) adapt to crisis all at once, or at different speeds for different parts? Closely linked parts changed almost instantly; the surrounding institutions turned out to be on their own centuries-long timeline unrelated to the crisis. The chapter also catches systems that *announce* change without actually changing underneath.

> *A note on what Part III is for. A reader could reasonably ask why a book about defining intelligence is about to spend two chapters on Dutch flood infrastructure and the way small children mislabel cows as dogs. The answer is that Part II built a piece of machinery — fidelity-weighted elimination, converging toward a hard core and a soft periphery — and machinery built in the abstract earns nothing until it has been pointed at something messy and watched to see what happens. These two chapters are not detours on the way to the real argument. They are the moment the book finds out whether its own formalism survives contact with cases it was not designed around. The first naive prediction tested here fails, instructively, in both domains. What survives the failure is more useful than what the naive version would have given if it had simply been confirmed.*

6.1 A clean prediction, stated in advance

Take a system with multiple coupled subsystems undergoing repair after a shared triggering failure, and ask: do they reach closure — do they stabilize a new, working response to the failure — at the same time, or at different times? If different times, is there a predictable pattern to the gap?

A natural first hypothesis, building directly on the active-inference-style alternative the fidelity framework was built to out-predict, would be that subsystems closer to the proximate cause of failure close first, and subsystems further from it — more socially or institutionally mediated — close later. Applied to a flood-control system with hydraulic, ecological, and institutional/governance subsystems, this predicts an ordering:

$$t_H < t_E < t_I$$

hydraulic recognition before ecological recognition before institutional reorganization, on

the reasoning that the physical failure is most immediate and the governance response most mediated and slowest.

This prediction is stated in advance, plainly, because what happens to it next is the chapter's actual content.

6.2 The case

The Netherlands' Room for the River program offers an unusually well-documented test. Severe floods on the Rhine and Meuse in 1993 and 1995 forced the evacuation of roughly a quarter million people, and the floods are widely credited as the starting point of a fundamental policy reconsideration — away from the centuries-old strategy of progressively raising and strengthening river embankments, toward a strategy of deliberately surrendering floodplain land back to the rivers. The full program, formally adopted via a 2006 Spatial Planning Key Decision, ran construction from 2007 to roughly 2015–2018 across more than thirty project sites.

The immediate post-flood response was not the eventual program. It was incremental: an adjustment of the design discharge standard the embankment system was built to handle, from 15,000 to 16,000 cubic meters per second — a parameter change within the existing engineering framework, not yet a structural reorganization. The decade between flood and formal structural decision is itself worth treating as data: it is roughly the gap between the triggering failure and the point at which the underlying engineering logic ("keep raising the embankments") was recognized, not as having failed once, but as having no terminus — a closure-exhaustion realization, arrived at only after the incremental route had visibly been tried and had visibly not resolved the deeper problem.

6.3 Where the ordering hypothesis breaks

The predicted ordering $t_H < t_E < t_I$ does not survive contact with the documentary record, and it breaks in two distinct and instructive ways.

First: hydraulic and ecological closure were not sequential at all. The program was explicitly designed, from early in its formation, around a joint objective balancing hydraulic effectiveness, ecological robustness, and spatial-cultural quality together — not ecological closure trailing years behind hydraulic closure, but the two being defined as a single combined admissibility criterion from the outset.

Definition 6.1 (Interface Lag)**.** For two coupled subsystems S_i, S_j with closure times t_i and t_j, define $\Delta t_{ij} = |t_i - t_j|$.

The ontology of subsystems

Before going further it is worth being explicit about what a "subsystem" is, since the word is about to carry real argumentative weight, here and even more so in Chapter 7, where

subsystem-indexed repair becomes the chapter's central repair move. Three questions are worth separating. Are subsystems discovered, existing in the world independently of anyone's analysis, or are they analyst-imposed, a partition the observer brings to the case rather than finds in it? Can subsystem boundaries change over time, or are they fixed once drawn? And what stops "subsystem" from becoming an infinitely flexible category, redrawn however convenient whenever a prediction needs rescuing?

This book's working position is deliberately modest rather than metaphysically ambitious: a subsystem is whatever a perspective (in the Chapter 3 sense) can independently measure and report on. Hydraulic engineering and institutional governance are treated as separate subsystems in this chapter not because reality is carved at uncontroversial joints, but because each has its own independent measurement methodology, its own historical record, and its own closure criterion that can be assessed without reference to the other — which is precisely the anti-vacuity standard this book holds subsystem-splitting to throughout, stated here first and then enforced explicitly as a named Principle when Chapter 7 needs it. On this view subsystem boundaries are not fixed for all time; they can shift as measurement capability shifts, which is itself informative rather than a defect — a boundary that only becomes visible once a new measurement technique exists is not arbitrary, it is discovered late. What rules out the worry that "subsystem" becomes infinitely flexible is exactly the discipline a reader should expect by now: no subsystem split is legitimate in this book without an independently motivated measurement method behind it, and chapters that introduce a split are expected to show that method, not merely assert the split.

Conjecture 6.2 (Coupling–Lag Relation)**.** Δt_{ij} is inversely related to the coupling strength C_{ij} between the subsystems — tightly coupled subsystems should reach closure near-simultaneously, while loosely coupled ones should drift apart.

The hydraulic–ecological pair in this case shows near-zero Δt_{HE}, consistent with high coupling — but this is not independent confirmation of the conjecture, since the two subsystems were defined together by program design rather than discovered to have converged. What the case actually offers here is an illustration of the high-coupling, low-lag corner of the conjecture, not a test of it: there was no opportunity for these two tracks to drift apart and then resynchronize, because they were never separated to begin with.

Second, and more substantively informative: the institutional/governance track does not share the 1993–1995 trigger at all. Dutch regional water-governance institutions — the canoe-house-era water boards, their consolidation from roughly 3,500 around 1900 to about 2,600 by 1950 to 21 today — trace a centuries-long reorganization driven by the growing power of national water-management authority since the late eighteenth century, a process already mid-stream when the 1990s floods occurred. Room for the River began as an ad hoc engineering reaction to a specific flood event and only afterward grew into application of a broader river-basin governance approach — meaning the institutional reorganization here was not delayed-response-to-a-shared-trigger at all. It was an independent, much older process the flood event happened to intersect and accelerate, not originate.

This breaks an assumption buried in the original hypothesis: that all subsystems share a common reference event from which a lag can even be measured. Where they don't, Δt_{ij} as defined is not well-formed — there is no single $t = 0$ from which both t_H and t_I can be measured, because the institutional clock was already running on its own schedule before the hydraulic crisis began.

6.4 A second, separate finding: nominal versus operational closure

A third phenomenon surfaces in the record, distinct from lag and worth a definition of its own. Official Dutch water policy doctrine proclaims water as the ordering principle for spatial planning — a stated, adopted closure at the level of policy language — yet actual spatial-development decisions are reported as continuing to follow prior economic and institutional priorities, with researchers concluding the resulting centralization tendency runs counter to what the new principles intended. The principle was formally adopted; the underlying practice it was meant to govern did not actually reorganize beneath it.

Definition 6.3 (Nominal vs. Operational Closure)**.** Let $N(t) \in \{0, 1\}$ denote whether a closure has been formally declared or adopted in a system at time t, and $O(t) \in \{0, 1\}$ denote whether the underlying practice has actually reorganized to match it. A *closure illusion* is a state in which $N(t) = 1$ while $O(t) = 0$ — the system has said it closed without having closed.

This is a different failure mode than lag. Lag is a question of timing between subsystems that do eventually both close. Closure illusion is a question of whether a declared closure is real at all, and it requires measuring N and O independently rather than treating a policy announcement as evidence of operational change — a methodological caution worth carrying into every later chapter that draws on institutional, cultural, or self-reported evidence of closure, since self-report is exactly where N and O are most easily conflated.

A related complication is worth naming rather than setting aside: several independent sources on flood infrastructure converge on the observation that structural hardening measures (levees, dams) tend to suppress the very signal that would otherwise prompt structural reorganization, because the public perceives risk as resolved and development on the protected floodplain increases as a result. Call this *patch-induced closure illusion* — a local fix that does not merely fail to trigger restructuring, but actively delays it by manufacturing the appearance that no restructuring is needed.

A brief illustration outside infrastructure

This concept is given a single, deliberately modest illustration from a third domain, offered as exactly that — one additional anecdotal case extending the concept's apparent reach, not a third leg of the formal two-domain comparison Chapter 8 builds and is careful to keep at $n = 2$. Shannon Vallor, Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence at

the Edinburgh Futures Institute, speaking at the "Is AI a Threat?" panel hosted by the Royal Institute of Philosophy on 26 June 2026, describes a finding from two independent surveys of executives (roughly two hundred UK chief executives in one study, three hundred US chief executives in another) using AI systems for business decisions: the large majority used AI for most of their decisions, trusted its advice over that of their human peers, and — in about half of cases — deferred to the AI's conclusion over their own judgment when the two conflicted. Vallor's own term for the end state of this pattern is *cognitive surrender*: a condition in which a decision-maker believes themselves to still be exercising judgment while the substantive work of recognition and implementation has, in practice, migrated to the system they are nominally only consulting.

The structural parallel to this chapter's central distinction is direct enough to be worth stating, while resisting the temptation to claim more than the single case supports. "Keeping a human in the loop" is, on this account, a nominal closure: $N(t) = 1$, an oversight mechanism is declared present. Whether $O(t) = 1$ — whether the human's presence is doing any actual epistemic work, rather than rubber-stamping a conclusion already reached elsewhere — is a separate, empirical question the declaration does not answer, and Vallor's CEO data suggests that for a substantial fraction of a real, surveyed population, it does not. Asked directly whether keeping a human in the loop addresses this risk, Vallor's answer was that the phrase functions, at best, as what she called a spongy guardrail, and at worst a total illusion, precisely because cognitive surrender is the condition of believing oneself still in the loop while no longer being so — a closure illusion stated in almost exactly this chapter's vocabulary, by a source that had no reason to be aware of it.

6.5 What this chapter actually establishes

Not the clean ordering it set out to test. What it establishes instead: tightly coupled subsystems can reach joint closure with negligible lag, consistent with — though not independently confirming — the coupling–lag conjecture; loosely coupled or independently-originated subsystems may share no common reference point at all, meaning lag is sometimes simply the wrong concept to reach for; and declared closure and operational closure are separate quantities that can diverge for extended periods, sometimes because a local patch actively obscures the divergence. These three findings recur, in a different costume, in the chapter that follows.

7

Development: Why the Clean Threshold Story Fails, and What Replaces It

In plain terms: A child's ability to *tell* two animals apart and their ability to *say* the right word for each turn out to run on separate, loosely connected tracks, often years apart — a toddler saying "doggie" for a cow may already perceptually distinguish the two. This forces tracking perception, comprehension, and speech as separate systems, with a strict rule against inventing such splits just to save a failing prediction.

7.1 The naive prediction, again stated in advance

A second domain, chosen for being about as unlike flood-control infrastructure as a case study can be: how children acquire and differentiate word categories. The naive hypothesis under test here is structurally identical to Chapter 6's, transplanted: a child maintains a single, unified category structure (a "distinction graph") that accumulates anomalies — instances that don't fit existing categories — until accumulated anomaly forces a closure-expanding reorganization, splitting an overextended category into two. Applied to the well-documented phenomenon of childhood overextension (a child calling cows, horses, and sheep all "dog"), the prediction is that new categories (dog vs. cow, as separate words) should emerge specifically when anomaly accumulation against the single existing category graph exceeds what the graph can absorb.

7.2 Where it breaks

This prediction does not survive contact with the developmental-psychology literature, and like the institutional case, it breaks in a way that is more useful than confirmation would have been.

First: production and comprehension dissociate. In preferential-looking studies, children showed no looking-time preference distinguishing "dog" from "cat" on anomalous trials (where no matching referent was present), yet did show a preference when "cow" was the requested label — and critically, prior overextension *in production* (the child saying "dog" for a cow) was not diagnostic of the child's underlying comprehension. A child can be overextending a word in speech while their comprehension already discriminates the categories that speech is failing to mark. This means there is no single "the child's category graph" whose saturation level can be tracked — there are at minimum two systems, production and comprehension, updating on different schedules and potentially containing different information at the same chronological age.

Second: in several documented cases, perceptual category differentiation precedes lexical differentiation rather than being generated by it. Eye-tracking work on infant categorization shows that visual category representations distinguishing cats from dogs, built from head-region cues, are established in early infancy — well before the one-to-two-and-a-half-year window in which overextension in speech is typically observed. The perceptual distinction (cat $\neq$ dog) is often already in place before the lexical failure (calling both "doggie") even begins; the lexical layer is not generating a new distinction under anomaly pressure so much as catching up to one that already existed in a different subsystem.

7.3 The repair

Definition 7.1 (Subsystem-Indexed Repair)**.** Rather than tracking a single global repair process $R_n(X)$ for "the child," index repair by subsystem: $R_n^{(k)}(X)$ for subsystem k — for example, perception, lexicon, or production.

This move is not free, and stating the cost of it plainly matters more here than almost anywhere else in the book, because it is exactly the kind of move that can quietly become unfalsifiable if not constrained. Splitting a global process into per-subsystem processes whenever a prediction fails is a standard and tempting way to rescue a theory from any disconfirming evidence whatsoever — simply declare a new subsystem boundary wherever the data doesn't fit, and no observation can ever count against the framework.

Principle 7.2 (Anti-Vacuity)**.** A subsystem split into index k is admissible only if there exists an independent measurement method M_k for that subsystem — established by existing methodology for reasons unrelated to rescuing this particular prediction, not introduced solely because the global version failed.

This is not a mathematical claim and has no proof; it is a methodological commitment, and it is satisfied here, not assumed: the perception/comprehension/production split in this chapter is not invented to save the anomaly-threshold hypothesis. It is borrowed

directly from existing developmental-psychology methodology — the preferential-looking paradigm exists, and was developed, precisely because researchers already suspected production and comprehension could dissociate, for reasons that have nothing to do with this book's elimination formalism. The split earns its use here because it was independently motivated before this chapter needed it.

7.4 A second, smaller dissociation

The within-subsystem story is also not clean. Even confined to comprehension and production separately, the literature warns against treating either as a single unitary process: the relevant developmental and self-regulatory literatures generally caution against assuming a shared core mechanism across what look, behaviorally, like related capacities — a theme that recurs with sharper formal content in Part IV, where the same warning becomes load-bearing for the recursive compression argument rather than a side note.

7.5 What survives

Not the clean anomaly-threshold hypothesis. What survives, and is consistent with — though again not independently confirmed by — Chapter 6's coupling-lag conjecture: subsystems within a single developing child can be loosely coupled enough to update on substantially different timescales, with perception sometimes leading production by a year or more, and apparent "category collapse" at one level (the spoken word "doggie" applied indiscriminately) can coexist with a fully maintained distinction at another (the eye-tracked perceptual discrimination between the two animals). This recasts what looked like a developmental error — overextension — as something closer to economical repair at the lexical/production interface specifically, with no failure at all occurring at the perceptual level.

8

What These Two Domains Together Establish

In plain terms: Two unrelated case studies produced the same underlying pattern — a tempting coincidence this chapter deliberately refuses to over-interpret. Two hand-picked cases are not statistical proof of anything general; they're a lead worth following, stated honestly as that and nothing more.

8.1 The structure that recurred

Two domains, chosen for having essentially nothing else in common — sovereign flood-control infrastructure and a toddler's vocabulary — produced the same qualitative pattern when the same naive ordering/threshold hypothesis was tested against each:

$$\mathcal{S} = (\Delta t,\ N/O)$$

— a joint structure combining measurable interface lag between coupled subsystems and a separate nominal-versus-operational closure distinction. Both domains showed: (a) at least one pair of subsystems with near-zero lag, consistent with high coupling; (b) at least one subsystem whose triggering event or originating timescale was independent of the others, breaking the assumption of a shared reference point; and (c) a distinction between declared/apparent closure and actual underlying reorganization, with the gap between them sometimes actively maintained by a mechanism that benefits from the appearance of resolution (the levee suppressing the signal for restructuring; production-level overextension persisting despite intact comprehension underneath it).

8.2 What this recurrence does and does not establish

It is tempting, having found the same structure twice, to claim the structure is general — a shared mechanism operating across institutional and cognitive-developmental systems

alike. That claim should be resisted in exactly the form it is tempting to state it.

Observation 8.1 (Structural Resemblance, not "Cross-Domain Invariance")**.** The same formal pattern $\mathcal{S}$ was observed in two domains chosen for reasons unrelated to testing this book's formalism — one institutional, one developmental — and was not the pattern either domain was originally selected to illustrate.

Why similar structures do not imply shared causes

It is worth stating explicitly, and not leaving to inference, exactly what kind of mistake the demotion above is designed to prevent — because the temptation it guards against is easy to fall into without noticing. Finding the same pattern in two domains licenses, at most, the inference *same pattern*; it does not license *same mechanism*. These are different claims, and the gap between them is not a technicality. Two systems can exhibit identical formal structure for entirely unrelated underlying reasons — a predator-prey population cycle and an electrical oscillator both solve the same class of differential equation, without either being a hidden cause of the other, or evidence that biology and electronics share a common substrate. Structural resemblance is consistent with shared mechanism, consistent with coincidence, and consistent with a deeper, currently unidentified common cause distinct from either domain studied — and nothing in finding the resemblance twice distinguishes among these three possibilities. What would distinguish them is a mechanism-level account connecting the two domains directly, which this book has not attempted and does not claim to have found. The honest content of this chapter's finding is exactly as stated above: a recurring formal pattern, motivating further investigation, supplying useful shared vocabulary for later chapters — and nothing about *why* the pattern recurs.

A natural-sounding but unjustified next step would assert something like $P(\mathcal{S} \mid D_1, D_2) > P(\mathcal{S} \mid D_1) \cdot P(\mathcal{S} \mid D_2)$ — that finding the pattern in both domains is itself evidence the domains share an underlying mechanism, by analogy to how correlated observations across independent samples support a common cause. This inference does not go through here. It presupposes a sampling process — domains drawn at random from some larger population of domains, with $\mathcal{S}$ then checked for co-occurrence — that does not exist in this book. Two domains were chosen, not sampled; and they were not selected entirely independently of each other, since the institutional case was investigated first and the developmental case was investigated partly because earlier work in this project had already developed vocabulary (subsystem indexing, interface lag) that made the developmental literature's production/comprehension dissociation easy to recognize once encountered. With $n = 2$ non-randomly-selected, non-independently-selected cases, no probability statement of this form is estimable, and stating one would borrow a statistical authority this evidence has not earned — precisely the error Chapter 5's machinery exists to catch, here applied to the book's own argument rather than to an external case.

What can honestly be said: the same distinction recurred in two domains selected for unrelated reasons, which motivates further investigation into whether it is a general fea-

ture of coupled repair systems, and supplies the book's working vocabulary (interface lag, coupling, nominal/operational closure) for the chapters that follow. It does not establish a shared mechanism. A reader entitled to remain unconvinced that anything general has been shown here is not making an error; they are correctly applying the standard this book has set for itself.

8.3 What carries forward

Three pieces of machinery, built and tested rather than merely asserted, carry into the rest of the book: the interface-lag concept and its open coupling conjecture; the nominal/operational closure distinction, which will recur explicitly in Part VI's discussion of borrowed authority, where a declared elimination and an operationally real one turn out to be exactly the same distinction in a different domain; and the anti-vacuity principle governing when a global process may legitimately be split into subsystems — a principle Part IV will need immediately, the first time this book asks whether "intelligence," like "the child's category graph," might also turn out to be more than one thing wearing a single name.

PART IV

The Pedagogy of Boundaries

9

Counterfoil Choice: A Cultural Technology for Training Admissibility Recognition

In plain terms: Among the Kalabari people, a person is sometimes offered two options where only one is ever meant to be chosen — the other exists purely to demonstrate what failure looks like. This chapter argues the device is a training contrast: making the wrong answer absurdly wrong lets a learner cheaply practice recognizing the *shape* of inadmissibility before facing subtler, higher-stakes cases.

> *This part exists to do something the previous one could not: find a domain where the elimination machinery of Part II is not merely tested against messy data, but where something resembling it is already, independently, a cultural technology — practiced, named, and reflected upon by the people who use it, centuries before this book's formalism existed to describe it.*

9.1 A tool the Kalabari already built

In 1999, Nimi Wariboko gave a name to something Kalabari elders had been doing for generations without needing to formalize it. He called it *counterfoil choice*: a structure in which two options are presented, but only one is a genuine alternative. The other — the counterfoil — is offered specifically to be refused. A father places a full bowl of garri and an empty bowl before his child. A bride is served an elaborate wedding feast and ritually turns her face from it. A young soldier is asked to throw his lame general at the enemy like a human cannonball. In each case, Wariboko writes, A is not an alternative to B; B is only a counterfoil to A. When an option is offered, the negative is put forward to show the person where the decision should not go. The decision-maker retains the formal freedom to choose either option, but the outcome is supposed to be foreclosed in advance — the freedom to choose is not limited the way it is in a Hobson's choice, yet the result of the selection resembles a Hobson's choice all the same.

Wariboko is explicit that this device is not, in his framing, a test of intelligence. He calls it a *culture wisdom quotient* test, designed to probe a person's internalization of what he terms the Kalabari aristocratic ideal — stylishness, nonchalant achievement, self-knowledge (*bu nimi*) — not their capacity to think. That disclaimer is the most useful sentence in the source material, because it draws a boundary this chapter needs and will respect rather than override: the question is not what the Kalabari say the practice tests. It is what the practice, examined as a piece of pedagogical engineering, actually trains — and the answer, this chapter will argue, is something prior to and more general than either intelligence or wisdom as ordinarily understood.

9.2 The counterfoil as a high-margin training contrast

A child does not learn the concept "edible" by being shown a single bowl of food and told its name. Wariboko's examples instead pair the admissible item with a deliberately, almost comically inadmissible one: a bucket against a basket that cannot hold water; a full bowl against an empty one. The distance between the admissible option and the counterfoil is not incidental. It is the entire pedagogical point — the lesson works precisely because that distance is made as large as possible.

Definition 9.1 (Counterfoil Pair)**.** A counterfoil pair is a tuple (a, c) where a is the admissible option, c is the inadmissible option, and the pair is selected so as to maximize $d(a, c)$, the distance between them, subject to remaining pedagogically relevant to the lesson at hand.

What is being installed by repeated exposure to such pairs is not the content of any single lesson — children do not need extensive practice to know a basket cannot hold water once the affordance is understood. What is being installed is something more general: the *shape* of a non-option. A child who has been shown enough wildly-mismatched pairs across enough domains — vessels, food, conduct — begins to acquire a transferable competence: the ability to recognize, quickly and with low cognitive cost, that something fails to satisfy a goal. The counterfoil is training wheels for a classifier, not a lesson about any particular bowl.

This reframes the formal object of inquiry. The interesting structure here is not a shrinking space of candidate definitions being narrowed by elimination, in the sense Part II built — that mapping was tried earlier in this project's development and abandoned as overreaching once it became clear the Kalabari material was not actually modeling definitional convergence. The interesting structure is a developing function

$$f : X \to \{0, 1\}$$

mapping candidate options to admissible or inadmissible, trained initially on pairs (a, c) chosen so that $d(a, c) \gg 0$, on the implicit pedagogical bet that competence built on exaggerated contrasts will eventually transfer to judgments made under much smaller margins, where the boundary is no longer obvious and must instead be inferred.

Analogy 9.2 (demoted from "Margin Theorem")**.** Counterfoil pedagogy resembles high-margin training in statistical classification, where larger separation between classes is associated empirically with faster or more robust acquisition of a decision boundary.

This is stated explicitly as an analogy and not a theorem: the formal guarantees underlying real margin-based learning results — assumptions about convexity, VC-dimension bounds, independent and identically distributed sampling — have no established analogue for how a human child acquires categories through cultural pedagogy. A genuine theorem here would require either a formal model of human concept acquisition under exaggerated contrast, which does not currently exist, or direct empirical evidence that children's learning rate scales with the size of a training contrast, which has not been cited in support of this chapter. The analogy is retained because it is evocative and organizes the intuition usefully — not because it has been proven to hold.

10

Two Axes: Recognition and Implementation

In plain terms: Knowing the right answer and actually doing it are not the same skill. This chapter formally separates *recognition* (telling right from wrong) from *implementation* (acting on it once recognized), and shows that most of what gets dismissed as foolishness in the Kalabari examples is really an implementation failure, not a recognition one.

10.1 A single margin obscures two different difficulties

The Kalabari material resists a single-margin story, and resisting it is what makes the material useful rather than merely illustrative. Consider two examples from the childhood (*awome*) stage of Wariboko's account. The bucket-and-basket test is difficult, if at all, only for a child who has not yet grasped basic physical affordances — once the affordance is understood, the choice is trivial, because nothing pulls the child toward the basket. Compare this to the drink-sharing ritual reported in the same source: the youngest male in a gathering is expected to serve the elders first and take only what remains, even when little is left and he plainly wants more. Here the difficulty is not perceptual at all. The child fully understands the rule. The challenge is resisting the pull of his own thirst.

These are different problems, and collapsing them into a single scalar margin obscures the difference. It is more accurate to track two quantities separately.

Definition 10.1 (Two-Axis Decomposition)**.** Let $d_p(a, c)$ denote the perceptual or conceptual distinguishability of a from c — how obvious the boundary is. Let $m(a, c)$ denote the motivational pull of the inadmissible option c, independent of how obvious the boundary is.

The bucket-basket case is high-d_p, near-zero-m: the boundary is glaring, and nothing tempts the child toward the wrong side of it. The drink-sharing case is near-zero-d_p, high-m: the boundary is perfectly clear, and the difficulty lies entirely in enacting it against one's own desire. The bibife marriage feast — offered to a grown bride who is, by the ethnographer's

own account, genuinely tempted by the aroma of food she knows is rightfully hers — is the adult-life extremity of the same axis: d_p effectively zero, m high and sustained over an entire ceremony.

Definition 10.2 (Recognition and Implementation Competence)**.**

$$I_R = P(f(x) = y)$$

— the probability of correctly classifying a candidate option as admissible or inadmissible —

$$I_I = P(\pi(x) = a \mid f(x) = a)$$

— the probability of correctly enacting the recognized-admissible choice, conditional on having recognized it correctly in the first place.

The conditional form of I_I is the important move here: by defining implementation success only over cases where recognition has already succeeded, the two quantities are cleanly separated by construction. A failure can now be diagnosed as belonging to one or the other, rather than collapsing into a single undifferentiated "made the wrong choice." Much of what gets called stupidity or failure across the Kalabari life-cycle examples is not a recognition failure at all. The bride knows exactly what she wants and exactly what is expected of her. The chief at his installation, asked to choose between a yam (self-provision) and a cannon ball (public defense), is not confused about which item the community demands he select. The soldiers who refuse to catapult their general are not uncertain about what their commander has, in his battle-fury, miscalculated. In each case recognition is not the bottleneck. Execution is.

Proposition 10.3 (Dissociation)**.** *High I_R does not imply high I_I.*

This is not given a formal proof here; it is supported constructively, by exhibiting cases — the adult-stage examples above — in which I_R is evidently near 1 while I_I is genuinely under strain. A fuller, independently-grounded version of this proposition is developed in Chapter 12, once the relevant psychological literature has been brought in.

Four failure modes

The conditional definition of I_I in terms of I_R makes it possible to lay out, exhaustively, every combination of high and low standing on the two axes, which gives the later transfer matrix of Chapter 12 an intuitive grounding before any matrix notation appears.

Recognition	Implementation	Interpretation
Low	Low	ignorance — neither the boundary nor its enactment is in place
High	Low	weakness — the boundary is seen clearly but not held to
Low	High	luck or habit — correct behavior without underlying understanding
High	High	competence — both recognized and reliably enacted

The middle two rows are the cells this book's recognition/implementation split was built to make visible, since a single undifferentiated notion of "success" or "failure" collapses them into indistinguishable outcomes. *Weakness* is the cell most of this chapter's Kalabari examples occupy — the bride, the chief, the soldiers all sit in this row, not the ignorance row a flatter theory of failure would place them in. *Luck or habit* is the more unsettling cell, easy to overlook: a person or system can behave correctly for an extended run without the recognition component ever having been present, which means correct behavior alone, observed from outside, cannot by itself distinguish this row from competence — a methodological caution worth carrying forward into Chapter 12's discussion of why transfer, rather than performance alone, is the test that actually discriminates between them.

A fifth case the table cannot represent

The table above assumes, without stating it, that I_R and I_I belong to the same agent being evaluated — that whatever recognized the boundary and whatever enacted it are the same locus of competence, even when they come apart from each other as the weakness and luck-or-habit rows describe. There is a fifth situation this assumption hides, worth naming explicitly rather than leaving for a reader to discover as an apparent exception to the table.

Shannon Vallor, in the panel discussion already cited in Chapter 6, describes a four-stage progression she terms *cognitive offloading*, *cognitive debt*, *cognitive deskilling*, and finally *cognitive surrender*: a sequence in which a person first lets an external system perform cognitive work on their behalf, then accumulates a growing gap between the work they could do unassisted and the work they have actually been practicing, then loses the underlying capacity such that performance degrades once the assisting system is removed, and finally arrives at a state of not being aware that the delegation has occurred at all — continuing to experience themselves as the one deciding, while the actual recognition and implementation have migrated elsewhere.

This describes a competence credited to an agent in the high-I_R, high-I_I cell of the table above, observed from outside, that does not belong there once the locus of recognition and implementation is correctly identified — because the agent credited with the competence is not, by the final stage of Vallor's cascade, the agent who actually possesses it. This is not

the same as the luck-or-habit row, where correct behavior occurs without recognition ever having been present in anyone; here recognition and implementation are both genuinely present, just not in the party the outcome gets attributed to. Call this *delegated competence*, distinct from all four cells above precisely because it requires relaxing the table's hidden single-agent assumption rather than moving within it. The unsettling part of Vallor's evidence, on her account, is not that this happens to novices or to populations conventionally treated as vulnerable, but that two independent surveys found it occurring in business leadership specifically — roughly half of surveyed chief executives, across both a UK and a US sample, reporting that they deferred to an AI system's conclusion over their own judgment when the two conflicted, a population an observer would, by performance alone, have every reason to place in the competence cell.

11

A Curriculum, Read Cautiously

In plain terms: The Kalabari life cycle looks, at first, like a tidy curriculum: easy lessons for children, hard willpower-testing ones for adults. A careful look at the early examples breaks that tidy story — willpower-testing lessons show up surprisingly early. The honest reading is a curriculum that blends both kinds of lesson from the start.

11.1 A pattern across the life cycle

Read across Wariboko's five life-cycle stages — *awome* (childhood), *asawo* (young adulthood), *opuasawo* (elders), *alapu* (chiefs) and their *iya* wives, and finally the status of ancestor — the examples suggest a developmental pattern. Early counterfoils are overwhelmingly high-d_p, low-m: the wrongness is visually self-evident and not particularly tempting. Counterfoils involving the apprentice trader's restricted commercial freedom, or the elder's restraint in administering punishment, occupy a middle zone where the boundary requires understanding of social and institutional structure rather than physical perception, and where real stakes begin to apply pressure. By the time one reaches chieftaincy installation, bibife, and the conduct expected of a dying person facing their final moments, d_p has fallen close to zero — no physical absurdity remains to lean on — while m remains high or increases further.

Definition 11.1 (Curriculum Trajectory)**.** A stage-indexed map $s \mapsto (d_p(s), m(s))$ tracking how perceptual margin and motivational pull change across an individual's developmental stages.

It is tempting to state this as a clean, monotonic curriculum: d_p falling and m rising in lockstep across five neat stages, with recognition training handed off cleanly to implementation training partway through. The temptation should be resisted, and resisting it here is an instance of exactly the discipline Part III's two case studies were built to model: a dozen well-chosen ethnographic anecdotes, selected by Wariboko to illustrate a different thesis (the aristocratic ideal), can be made to fit more than one cur-

riculum shape, and treating any one shape as established by the examples rather than merely consistent with them would repeat the error both Chapter 6 and Chapter 7 were built to diagnose and correct.

Closer inspection of the *awome*-stage material itself undermines strict sequentiality: the drink-sharing and shared-plate rituals are pure implementation challenges (low d_p, real m) embedded early, well before a clean handoff point would predict. The more defensible reading is interleaved rather than sequential.

Conjecture 11.2 (Interleaved Exposure)**.** Rather than a strict handoff $I_R \to I_I$, effective training exposure at stage s is better modeled as a weighted combination,

$$E_s = \alpha_s I_R + \beta_s I_I,$$

with $\alpha_s > \beta_s$ early in development, and the ratio shifting toward β_s as the individual matures.

This is stated as a conjecture, not a theorem, and not even a fully established proposition: it is motivated by, but not established by, the same dozen non-randomly-sampled illustrative cases just discussed. The anti-vacuity principle introduced in Chapter 7 applies here directly and is worth restating explicitly rather than left implicit: this curriculum hypothesis is admissible to entertain because the recognition/implementation distinction it relies on was independently motivated in Chapter 10, before this curriculum question was asked — not invented here to make Wariboko's examples fit a pre-decided shape.

12

Does It Transfer? — The Four-Cell Question

In plain terms: If recognition and implementation really are separate skills, does practicing one help the other? Neither flat independence nor full unity survives the evidence. The two are clearly distinguishable but not cleanly separate — entangled, not identical.

12.1 Why transfer is a stronger test than performance

Before the matrix machinery, it is worth being explicit about why this chapter asks about transfer at all rather than stopping at performance, since the move from one to the other is not obvious to a reader encountering it for the first time. A competence that only succeeds in the exact environment it was trained or observed in is weaker evidence of a real, general capacity than one that succeeds somewhere new — because the first case leaves open a deflationary explanation the second case rules out. A child who reliably refuses the basket-for-water counterfoil might be pattern-matching a specific, memorized scenario rather than possessing any general competence in admissibility recognition; a child who recognizes an unfamiliar, never-before-seen counterfoil using the same underlying logic is harder to explain away this way. Performance in the training environment is therefore necessary but not sufficient evidence for the kind of competence this book is interested in tracking. Transfer — performance somewhere the training did not directly prepare for — is the test that actually discriminates a genuine, generalizable capacity (the *competence* row of the previous chapter's table) from the case where correct behavior was produced some other way (the *luck or habit* row). This is why the four-cell question below matters more than it might first appear: it is not an incidental refinement of the recognition/implementation split, it is the test that determines whether the split was tracking something real in the first place.

12.2 The transfer matrix

If recognition and implementation are genuinely distinct competencies, the natural next question is whether training in one transfers to performance in the other, and the question splits into four parts rather than one.

Definition 12.1 (Transfer Matrix)**.**

$$T = \begin{pmatrix} T_{RR} & T_{RI} \\ T_{IR} & T_{II} \end{pmatrix}$$

where T_{RR} measures whether recognition practice at high margins improves recognition at lower margins; T_{II} measures whether implementation practice under one temptation improves implementation under another; T_{IR} measures whether implementation practice improves later recognition, particularly recognition under pressure; and T_{RI} measures whether recognition practice improves later implementation.

A first pass through this question risks an overly tidy answer: strong within-faculty transfer ($T_{RR}, T_{II} > 0$) alongside near-zero cross-faculty transfer ($T_{RI} \approx T_{IR} \approx 0$), which would license a clean story of two largely independent developmental pathways — the familiar folk distinction between being "brilliant but self-destructive" versus "wise but ineffective," reframed as two points in an (I_R, I_I) plane rather than as a contradiction.

Proposition 12.2 (Orthogonality Failure)**.** *If $T_{IR} > 0$ or $T_{RI} > 0$, then I_R and I_I are not orthogonal in the relevant sense.*

This is true essentially by the definition of what nonzero cross-transfer means, and is stated for clarity rather than as a result requiring independent proof. Whether the antecedent holds — whether T_{IR} or T_{RI} is actually nonzero — is the empirical question this chapter exists to address.

12.3 What the literature shows

The clean independence story does not survive contact with the developmental and social-psychology literature, and it fails in an informative rather than merely disappointing way. The construct of *moral disengagement* — the use of self-monitoring, judgment, and self-reactive mechanisms to permit wrongdoing one already recognizes as wrong — maps closely onto the high-I_R, low-I_I cell, and its existence as an independently studied construct is real evidence that recognition and implementation dissociate. But a longitudinal finding complicates flat independence in the other direction: self-control and cooperative behavior in childhood predict reduced moral disengagement in adolescence — early implementation-relevant

practice associated with a later, partly recognition-adjacent outcome. This is correlational rather than experimental evidence, and the direction of any underlying causal mechanism should be stated as a hypothesis the finding is consistent with, not a demonstrated mechanism — but it is a documented nonzero association, inconsistent with a flat $T_{IR} \approx 0$ claim. Separately, the broader self-regulation literature warns against treating I_I as a single transferable competence at all: executive function, emotion regulation, response inhibition, and effortful control are measurably distinct, overlapping but non-identical constructs, which means even T_{II} — transfer within implementation alone — should not be assumed to hold across arbitrarily different forms of temptation.

Proposition 12.3 (Dissociability)**.** *If there exist population subsets in which $I_R \gg I_I$, and other subsets in which $I_I \gg I_R$, then the two capacities are non-identical.*

Also near-definitional — the existence of such subsets is itself the content of the claim — and supported here by citing literature consistent with both subset types existing (moral disengagement for the first; documented cases of strong impulse control without corresponding insight for the second) rather than derived mathematically.

12.4 The honest formal conclusion

$$I_R \not\equiv I_I \qquad \text{but} \qquad I_R \not\perp I_I$$

Recognition and implementation are dissociable as constructs, not identical — but they are not cleanly independent either. They are partially coupled, developmentally entangled in at least one demonstrated direction, and — as the self-regulation literature insists — individually multidimensional rather than scalar. This last point reframes I_I itself:

$$I_I = (I_{\text{inhibition}}, I_{\text{emotion}}, I_{\text{effort}}, I_{\text{delay}}, I_{\text{social-cost}}, \ldots)$$

a vector of partially correlated sub-competencies rather than a single number. Implementation, offered in Chapter 10 as the second half of a clean two-term split, turns out under inspection to be its own compressed bundle. This is the observation Chapter 13 exists to take seriously rather than treat as an inconvenience.

13

The Recursion

In plain terms: Intelligence turned out to be two things wearing one name; then implementation itself turned out to be several things wearing one name. The chapter's claim: this is what happens, in general, whenever a culture gives a single word to a cluster of related but separable skills. It also corrects a real mathematical error from an earlier draft, where "compression" was wrongly treated as reversible.

13.1 A complication that turns into a principle

The same operation that split intelligence into (I_R, I_I) has just been applied again, inside I_I itself, and has produced the same shape: a single named competence dissolving, under inspection, into a partially correlated bundle of lower-level competencies. This is not a complication to be managed. It is the chapter's actual claim.

Definition 13.1 (Recoverable Compound)**.** A competence label ℓ is *recoverable-as-compound* if there exist measurable variables $(x_1, \dots, x_n)$, with $n > 1$, and an aggregation map $g : \mathbb{R}^n \to \mathbb{R}$ such that $\ell \approx g(x_1, \dots, x_n)$, and the x_i are not mutually redundant — no x_i is a deterministic function of the others.

This definition deliberately replaces an earlier, mathematically unsound formulation considered during this project's development, in which a compression operator $C : \mathbb{R}^n \to \mathbb{R}$ was paired with a claimed inverse C^{-1} recovering the original components. A many-to-one map has no general inverse; that earlier formulation was not a coherent mathematical statement, however suggestive it sounded. What the chapter's actual evidence supports is empirical recoverability of correlated sub-components — $I_{\text{inhibition}}$, I_{emotion}, I_{effort}, and so on are independently measurable and only partially correlated with one another — not deduction of those components from the compressed label alone. The corrected definition states only what has actually been shown.

Conjecture 13.2 (Recursive Compression)**.** Whenever a culture names a competence with a single word, closer inspection tends to reveal that the label is recoverable-as-compound in the sense just defined.

This is stated precisely and is, in principle, testable against further cases — and it currently rests on exactly one worked domain. Intelligence was split into (I_R, I_I); I_I was, on inspection, split again. That is one chain of decompositions, traced once, not yet a demonstrated general law. Whether it generalizes — whether this is something true of competence-words broadly, or an artifact specific to intelligence — is a question this chapter does not answer and should not be read as having answered. It is the question Part V exists to test, using a deliberately different competence word, chosen specifically because it offers an independent route to the same kind of evidence this part has relied on rather than merely repeating the method on more intelligence-adjacent material.

Compression is not reduction

The recursive compression principle can easily be misread as a form of reductionism, and it is worth heading off that misreading directly rather than letting it stand. It is not a reductionist claim, and the difference matters.

When a label ℓ compresses a bundle $(x_1, \ldots, x_n)$, the claim $\ell \approx g(x_1, \ldots, x_n)$ says that the label allows observers to coordinate around the bundle without repeatedly enumerating its components. It does not say the components cease to exist, or that they are somehow less real than the label that summarizes them. A map compresses a territory; it does not destroy the territory. A library catalog compresses a collection of books into a small number of searchable fields; it does not eliminate the books, or imply that a book is "nothing but" its catalog entry. A competence label behaves the same way: it compresses a bundle of underlying capacities for the convenience of everyday reference, without implying that the bundle is somehow less structured, less real, or less worth investigating than the single word used to refer to it in casual conversation.

This distinction matters specifically because the word "intelligence" is so often treated as though it names a single hidden essence waiting to be correctly identified — the discovery-cycle assumption this book opened by questioning. The recursive compression principle proposes something weaker and, on reflection, more ordinary: intelligence functions as a useful summary description of many partially independent capacities, the way most everyday competence words do. The compressed label is epistemically downstream of the structure it summarizes, not a substitute for investigating that structure.

Recoverability, in the sense already given a precise definition above, is the test that distinguishes a genuine compression from a label that has stopped functioning as one. A successful compression permits partial reconstruction of its underlying components — which is exactly what this book's Part IV case study did, in showing that I_R and I_I are independently measurable beneath the single word "intelligence." If no such reconstruction were possible even in principle, the label would not be compressing anything; it would simply be a name, with no underlying structure for "compression" to mean anything about. The purpose of the recursive analysis pursued across this book is therefore not to reduce intelligence to simpler, more fundamental parts in the way reductionist programs typically aim to. It is to identify

the structure that a convenient single word has been quietly summarizing all along.

Compression creates blind spots

The previous subsection's point — that compression does not destroy the structure it summarizes — should not be read as implying the compression is therefore costless. It is not. Every compression $g(x_1, \dots, x_n) \to \ell$ necessarily loses information in the specific sense that ℓ alone, without independent access to g and the x_i, does not in general allow that information to be recovered; the territory survives the map, but anyone holding only the map cannot see the territory's full detail from the map alone. This is the sense in which the convenience purchased by a compressed label is inseparable from the distinctions it hides: the same operation that makes "intelligence" usable in a sentence without enumerating five sub-competencies is the operation that makes a listener, hearing only the word, unable to tell from the word alone which of those five sub-competencies is actually in play in a given case.

This is not a flaw specific to the word "intelligence," or even to competence-words generally; it is a structural property of compression as such, and it recurs wherever a single label stands in for a more differentiated reality — political and ideological labels that compress a candidate's actual position on dozens of separable issues into one word a voter can hold in mind; educational labels ("gifted," "struggling") that compress a profile of distinct strengths and weaknesses into a single placement; and, not least, the discovery cycle that opened this book, where each generation's confident definition of intelligence was itself a compression that hid, from the people using it, exactly the substructure later generations would need to rediscover before the definition could be repaired. Naming this explicitly matters because it reframes what kind of carelessness the discovery cycle actually was: not a series of people failing to look hard enough, but a series of people working, unavoidably, at the resolution their compressed vocabulary allowed them to work at, until something — a new case, a new measurement, a new perspective — forced the blind spot into view.

13.2 An asymmetry worth naming rather than hiding

One more thing should be said plainly before this part closes, because it is the kind of gap a careful reader will notice and an honest book should flag first. I_R has, throughout this part, been treated as if it resists the fragmentation that just happened to I_I — recognition has stayed a single scalar quantity, $P(f(x) = y)$, while implementation fractured into a vector. There is no principled reason, internal to anything argued here, why recognition should be exempt from the same recursive move. It may simply be that this part's evidence happened to test implementation's internal structure and never tested recognition's. If the recursive compression conjecture is right, I_R itself is a plausible next candidate for decomposition — perceptual discrimination, social-context recognition, rule-recall, and pattern-completion may turn out to be as separable from one another as inhibition is from delay-tolerance. That

this part stops one level short of testing this is a limitation of what has been done here, not a finding that recognition is somehow special.

PART V

A Second Pass

14

Creativity: A Second Test of the Recursive Compression Principle

In plain terms: The recursive compression idea was built on one word, intelligence — not enough to call it a general law. This chapter tests it against creativity, and finds a partial match: the recognition/implementation split recurs, but creativity needs extra pieces intelligence didn't. A partial match is more convincing than a perfect one would have been.

> *Chapters 9 through 13 showed that the recursive compression machinery can organize one case — intelligence, examined through the Kalabari material and the recognition/implementation literature — with internal coherence. That is not yet evidence the machinery generalizes. A framework built around a single domain can always be made to fit that domain; the only way to find out whether it is tracking something real about competence-words in general, rather than something specific to intelligence, is to point it at a second word chosen for reasons unrelated to making the first case look good, and see what happens. This part is that attempt, and unlike Parts III and IV, it does not end with a clean repaired model. It ends with an open question, stated as precisely as this book can currently state it, and left that way rather than dressed up as more finished than it is.*

14.1 Why this chapter is not really about creativity

Chapter 13 closed with a conjecture and an admitted asymmetry: the recursive compression principle had been demonstrated once, on one word, decomposing in one direction (implementation fractured; recognition was never tested). Without a second case, that conjecture is indistinguishable from a description of something specific to intelligence — perhaps intelligence is unusually bundle-prone because of its particular evaluative stakes or its role in social sorting, and no other competence word behaves this way at all. This chapter exists to find out, and the honest framing of what it is actually testing matters more than the specific word chosen to test it.

Definition 14.1 (Model Transport)**.** Let a decomposition framework — here, the recognition/implementation split and the recursive compression conjecture built around it — be

said to *transport* to a second domain if a structurally analogous decomposition can be independently motivated in that domain, using evidence not originally assembled to support the first domain's case.

The question this chapter asks is not "is creativity also complicated?" — every serious treatment of creativity already agrees that it is, and that agreement by itself would prove nothing about whether this book's specific machinery is tracking something real. The question is narrower and sharper: does creativity decompose along a structurally similar axis to the one intelligence decomposed along — something recognition-like (the capacity to identify what counts as a good or admissible candidate) and something implementation-like (the capacity to actually produce and follow through on it) — or does it instead fracture along some entirely different axis that this book's machinery would not have predicted? Either answer is informative. Only the first counts as a successful transport.

14.2 Creativity, examined on its own terms

The existing literature on creativity, developed independently of anything in this book and for entirely different reasons, already resists treating creativity as a single faculty. Componential and systems-level accounts of creativity — independently of this book's formalism and predating it by decades — typically distinguish at least: divergent generation (the capacity to produce many candidate ideas, often measured by fluency and flexibility on open-ended tasks); evaluative or convergent judgment (the capacity to assess which of many generated candidates is actually good, original, or worth pursuing — a distinct skill from generating candidates in the first place, since fluent generation and good judgment are not the same trait and do not always co-occur in the same individual); domain expertise (creative output in most serious domains depends heavily on accumulated technical knowledge of that domain's existing constraints and possibilities, which is acquired separately from either generation or evaluation); persistence and follow-through (the capacity to carry a generated and evaluated idea through the often long, unglamorous work of actual production, revision, and completion, as distinct from having had the idea); and social or field-level uptake (whether a domain's existing experts and institutions actually recognize and validate the output as creative, which depends on factors external to the individual entirely).

This is not new evidence assembled by this book to support its own thesis. It is the existing shape of how creativity researchers, working independently and for independent reasons, already describe the construct. What is new here is asking whether this independently-arrived-at decomposition maps onto something recognition/implementation-shaped.

Definition 14.2 (Structural Isomorphism Test)**.** A map φ from intelligence's decomposition to creativity's decomposition preserves the relevant structure if $\varphi(I_R)$ corresponds to a recognition-type component of creativity and $\varphi(I_I)$ corresponds to an implementation-type component, with the correspondence doing real classificatory work — that is, with most of

the literature's named creativity components falling clearly on one side or the other rather than being ambiguous or evenly split.

Applying this test honestly: divergent generation and evaluative judgment both sit closer to I_R — they are forms of recognizing or producing candidates against some standard of admissibility, even though generation and evaluation are not the same skill and may themselves eventually warrant their own further split, exactly the kind of within-faculty fragmentation Chapter 13 found inside I_I. Persistence and follow-through sit clearly on the I_I side — they are about enacting a recognized-good idea against the resistance of difficulty, boredom, and the temptation to abandon unfinished work, structurally close to implementation-under-pull as defined in Chapter 10. Domain expertise and social uptake do not map cleanly onto either side, and this is worth stating plainly rather than forcing a fit: they look like separate axes the intelligence case never had to contend with, since nothing in the Kalabari material involved an external community certifying whether a choice counted as intelligent the way a creative field certifies whether a work counts as creative.

Conjecture 14.3 (Generalization, partially supported)**.** The recursive compression principle transports to creativity in part: a recognition/implementation-like split is independently recoverable from existing creativity research, supporting the conjecture's core claim that competence words generally conceal a recognition-implementation-shaped internal structure. But creativity's decomposition is not a clean copy of intelligence's — it carries at least two additional axes (domain expertise, social uptake) that intelligence's worked example in this book did not surface, which suggests either that intelligence has a parallel hidden structure this book's Part IV case study simply never tested for, or that the recursive compression conjecture is real but the specific recognition/implementation framing is only one instance of a more general decomposition pattern, not the pattern's full content.

14.3 What this chapter does and does not establish

This is reported honestly as partial confirmation with a genuine complication, not as a clean second success. The recursive compression conjecture does not appear to be an artifact specific to intelligence — that much transports. But the specific two-axis (I_R, I_I) shape does not transport whole; creativity needed more axes than intelligence's worked example used. The more defensible conclusion, and the one this chapter actually earns, is narrower than either a triumphant confirmation or a clean failure:

$$\text{Intelligence} \neq \text{Unique Case}.$$

Competence words generally seem to compress more structure than a single label suggests, and recognition/implementation is one recurring fault line along which that structure splits — evidently not the only one, and this book has not characterized what determines, for a given competence word, how many axes its internal structure actually has.

15

Why Compress at All?

In plain terms: If ”intelligence” and ”creativity” hide so much structure, why hasn’t language already replaced them with more precise vocabulary? The candidate answer: single words survive because they’re communicatively *efficient*, not because they’re correct. This chapter states that idea precisely and is honest that it doesn’t yet prove it.

15.1 The objection this chapter exists to answer

A reader who has followed this book to this point has watched intelligence decompose into recognition and implementation, implementation decompose further into a vector of inhibition, emotion regulation, effort, delay, and social cost, and creativity decompose along an overlapping but not identical set of axes. The natural and entirely fair objection at this point is: if these single-word labels are this inaccurate — if “intelligence” and “creativity” are each concealing a half-dozen only-partially-correlated sub-competencies — why do cultures keep using single words for them at all? Why hasn’t ordinary language already fragmented into the more accurate vocabulary this book keeps reaching for?

This chapter does not fully answer that question. It states the question precisely enough to be answered, sketches the shape an answer would need to take, and is explicit about stopping short of supplying one. That is the honest status of the material here, and the chapter is better for saying so directly than for dressing an unfinished argument in the language of a completed one.

15.2 A candidate account, stated as an open program rather than a result

Conjecture 15.1 (Compression Efficiency)**.** A competence label persists in a language when, and roughly because, the communicative savings of using one word rather than a full multidimensional profile exceed the predictive error that compression introduces.

Stating this precisely enough to be more than a slogan requires two quantities this chapter can define but not yet specify with any actual content.

Definition 15.2 (Description Length and Explanatory Error)**.** Let $L(n)$ denote the communicative cost — in whatever unit eventually proves appropriate, plausibly something like processing time, working-memory load, or message length — of describing a competence using n independent dimensions rather than a single compressed label. Let $E(n)$ denote the predictive or explanatory error that remains when a competence is described using only n dimensions rather than its full underlying structure — the gap between what an n-dimensional description predicts about a person's behavior and what actually happens.

On this account, a single-word label is communicatively efficient exactly where collapsing from n dimensions to 1 saves more in L than it costs in E:

$$\Delta L \gg \Delta E \quad (\text{at } n = 1, \text{ relative to the full decomposition}).$$

If something like this inequality holds for "intelligence" and "creativity" as actually used, it would explain why decomposition keeps happening in careful analysis — researchers, ethnographers, and this book itself keep finding more structure underneath the label — while the vocabulary itself remains stable in ordinary use: the label is not false, on this account, so much as it is a genuinely efficient lossy compression, one that most everyday uses of the word do not need to undo, because most everyday predictive purposes (deciding roughly whom to trust with a difficult task, whom to recommend for a job, whom a child should try to emulate) tolerate the error $E(1)$ introduces in exchange for the enormous savings in L.

The folk phrases this book has returned to several times — "brilliant but self-destructive," "wise but ineffective" — are, on this account, exactly the moments where the compression's error becomes locally too costly to ignore, and language responds by reaching for a longer, more decomposed description precisely because the single-word version has stopped being efficient for the purpose at hand. That the longer phrase exists at all, rather than language simply tolerating a wrong prediction, is itself a small piece of evidence that ordinary language is doing something like cost-sensitive compression rather than blind compression.

Why not just keep every distinction?

A sharper version of this chapter's question is worth asking directly: if compression genuinely costs explanatory accuracy, why does language compress at all, rather than simply maintaining an unbounded, maximally precise vocabulary — a separate word for every distinguishable combination of recognition and implementation sub-competencies, with nothing ever collapsed into a coarser label?

The candidate answer available from the machinery already built in this chapter is that an infinite vocabulary, even if cognitively achievable, would maximize precision at the direct expense of coordination. Communication requires that a speaker's word and a listener's understanding of that word refer to approximately the same thing without extensive negotiation on each use; the more finely a vocabulary is allowed to fragment, the less likely any

two speakers share the exact same fine-grained term for the exact same fine-grained bundle, and the more communicative work — exactly the cost $L(n)$ this chapter has already introduced — is required merely to establish shared reference before any actual content can be exchanged. A finite vocabulary trades away some precision in order to purchase a public, shared coordination surface: a word that is somewhat lossy but reliably understood the same way by most speakers most of the time is more useful for the actual business language is for — coordinating action and belief across many speakers — than a maximally precise word understood by no one but its coiner.

This connects directly to a theme that recurs elsewhere in work adjacent to this book, on why some distinctions persist across long stretches of time while others are quietly let go: a distinction survives in a shared vocabulary not simply because it is real or because it could in principle be drawn, but because it continues to earn its communicative keep — because enough speakers, enough of the time, need to coordinate around exactly that boundary for the cost of maintaining a dedicated word for it to be worth paying. An infinite vocabulary would maximize the precision available to a single isolated mind. A finite one is what a vocabulary used by more than one mind, repeatedly, under real time pressure, actually evolves toward.

15.3 What is missing, named explicitly

This is as far as the conjecture can honestly be taken in this book, and what is missing is not a minor gap to be waved past. L and E have been defined in words, not specified with any actual content — there is no stated unit in which L is measured, no model of what E actually computes, and no derivation showing the inequality $\Delta L \gg \Delta E$ holds for any real competence word rather than merely sounding plausible when described in prose. This is, structurally, the same kind of gap Chapter 5 left open with the Misweighted Elimination conjecture: a precisely statable, falsifiable-in-principle claim that this book has not closed, offered honestly as the chapter's research target rather than smuggled past as an already-finished result.

This book ends, on this question, not with a theorem but with a research program: a serious, well-motivated conjecture — that competence words survive not because they are fundamental but because they occupy efficient positions on a communication-cost-versus-explanatory-accuracy frontier — stated precisely enough that someone, possibly the author of a later book, could build the cost model this chapter does not, and either close the gap or discover that the inequality fails and a different account of why competence words persist is needed instead.

15.4 What this completes, and what it leaves open

Part V was undertaken to find out whether the machinery built in Parts II through IV survives contact with a domain it was not built around. Chapter 14 found a genuine, if partial,

transport: creativity decomposes along an overlapping but not identical axis to intelligence, supporting the recursive compression conjecture's core claim while complicating its specific two-axis shape. This chapter found that the deeper question the whole program eventually raises — why compress at all, if compression loses this much — is real, answerable in principle, and not yet answered here. Both outcomes are reported as what they are. The book attempted its stress test. It did not fully resolve what the stress test surfaced. What survives the attempt — Intelligence $\neq$ Unique Case, and a precisely stated open conjecture about why competence words persist despite the inaccuracy this book has spent thirteen chapters documenting — is offered as exactly that much, no more, and the next chapter's resolution of the discovery/constitution fork does not depend on any more than that having been shown.

PART VI

Resolving the Fork

16

Bad Counterfoils: A General Theory of Borrowed Authority

In plain terms: One idea from Chapter 5 — that mistakes happen when low-quality evidence gets treated as high-quality — turns out to explain failures everywhere: institutions that declare success without changing, moral panics, exaggerated warnings. This chapter states that pattern directly as the book's single most exportable idea, useful even to a reader who rejects everything else here.

> *A note on sequence. This part was drafted with full knowledge of Part V's actual outcome (Chapters 14–15: a partial, not clean, transport of the recursive compression conjecture to creativity, and an explicitly unfinished compression-efficiency conjecture). That outcome is referenced directly below where relevant. The point made here is not that Part V's result was unknown — it is that this part's central argument does not depend on which way Part V came out. The argument was already complete by the end of Part II, and Part V's partial, imperfect result is, if anything, a better test of that independence than a clean success would have been: the reader can check, concretely, that Chapter 17's conclusion does not lean on Part V having gone better than it did.*

16.1 One sentence doing more work than it looked like

Chapter 5 produced a sentence that was supposed to be a summary, not a discovery in its own right: intelligence-as-self-description fails when low-fidelity counterfoils are mistaken for hard eliminations. Across the chapters since, that sentence has kept reappearing under different names. Chapter 6 found institutions declaring closure ($N(t) = 1$) without having operationally achieved it ($O(t) = 0$) — a nominal elimination mistaken for a real one. Chapter 7 found a developmental literature in which production-level overextension looked, from the outside, like a recognition failure, when comprehension underneath it was already intact — an apparent boundary mistaken for a real one. Chapter 9 found an entire cultural pedagogy organized around manufacturing low-stakes, high-margin counterfoils precisely so a learner could practice telling forced eliminations apart from real ones before the stakes got serious. The pattern was never confined to intelligence. This chapter states it on its own terms.

16.2 The general theory

Definition 16.1 (Borrowed Authority)**.** Let q_i be the source-independence component of fidelity introduced in Chapter 5 — the degree to which an elimination's authority survives being stripped of the social standing of whoever asserts it. Define

$$B_i = 1 - q_i,$$

the *borrowed-authority score* of constraint i: how much of its apparent force depends on the teller rather than the world.

Definition 16.2 (Counterfoil Distortion Index)**.** Let D_i be the ratio of an eliminating claim's depicted severity to its actual severity:

$$D_i = \frac{\text{depicted severity}}{\text{actual severity}}.$$

A large D_i indicates rhetorical amplification — the claim is dramatizing an outcome beyond what the evidence supports, which corresponds to low severity calibration (s_i, from Chapter 5) for the same constraint.

Proposition 16.3 (checked for circularity)**.** *World-forced eliminations are characterized by high overall fidelity w_i; rhetorical eliminations are characterized specifically by high B_i.*

Stated this baldly, the proposition risks being empty rather than informative, and it is worth being explicit about exactly where the emptiness would come from rather than letting a reader discover it unaided. Since q_i is one of the four multiplicative factors composing w_i ($w_i = d_i \cdot s_i \cdot c_i \cdot q_i$, in the notation fixed in Chapter 5), an elimination with low q_i will tend to have low w_i as well, simply by the arithmetic of the formula — unless the other three factors happen to compensate. "Rhetorical eliminations have high B and low w" is, read this way, close to a restatement of how B and w are defined in terms of the same underlying q, not an independently derived empirical finding.

The chapter's actual, non-circular claim is narrower and more interesting than the proposition as first stated, and it is the claim worth defending: in the cases examined across this book — the empty-bowl pedagogy of Chapter 9, the rhetorically amplified institutional myths this chapter goes on to discuss — low q_i tends to co-occur with low d_i, low s_i, and low c_i *simultaneously*, rather than independently. The empty bowl is not merely authority-dependent (low q); it is also low-directness (poverty does not follow deterministically from a single act of childhood laziness), low-severity-calibration (the depicted certainty of ruin vastly overstates the actual conditional probability), and low-closure (real recovery paths — charity, family support, second chances — exist and are simply not shown). This joint co-occurrence across all four components is what the chapter claims, and it is an empirical pattern about how low-fidelity counterfoils are actually constructed in practice, not a logical consequence of how the four components happen to be defined. It is also, importantly, not established

with statistical rigor here — it rests on the worked examples examined in Chapter 9 and the case sketched below, and should be read with the same caution Chapter 8 applied to its own two-domain recurrence: suggestive, not demonstrated.

16.3 Exporting the machinery

Political ideologies, religious systems, moral panics, risk communication, and institutional self-justification all run on collections of counterfoils — exaggerated, rhetorically amplified non-options offered to make a predetermined conclusion feel self-evident, structurally indistinguishable in form from the empty bowl, differing only in the domain and the stakes.

Take a single brief case to show the machinery travels rather than merely asserting that it does. A moral panic — the recurring pattern in which a specific, often statistically rare harm is presented as an imminent, near-certain threat to children or to social order, justifying urgent restriction of some named activity — can be scored against all four fidelity components directly. Causal directness is typically low: the panic's narrative usually requires several unstated, contingent intervening steps between the named activity and the depicted harm. Severity calibration is typically low: D_i, the ratio of depicted to actual severity, tends to run far above 1, since panics are, definitionally, panics partly because the depicted risk outpaces the measured one. Closure is typically low: panics rarely acknowledge the recovery paths, safeguards, or base-rate context that would weaken their force. And source independence is typically low: panic-narratives draw heavily on the authority of the messenger (an institution, a movement, a media figure) rather than on claims that would retain their force if asserted by an unaffiliated stranger. The same four-component audit that distinguished the stove warning from the empty bowl in Chapter 9 distinguishes a documented public-health risk communication from a moral panic dressed in its vocabulary — and does so by the same mechanism, applied outside intelligence entirely.

A second, more contested case: deflection

A second case is worth including specifically because it is harder than the first, and the difficulty is instructive about a limit of this chapter's machinery rather than a flaw in it. At the same "Is AI a Threat?" panel already cited in Chapters 6 and 10, panelist Jens Munch, a technology and AI strategist, entrepreneur, and investor, argued that public and policy attention to speculative risks from rogue superintelligence or artificial general intelligence functions as a convenient deflection from a more mundane, more immediately tractable problem: as AI systems are continuously updated and re-versioned, legal liability for their decisions becomes structurally difficult to assign, eroding corporate accountability and democratic oversight in ways current legal frameworks are not equipped to track.

The structure of this claim is exactly the structure Chapter 9 introduced and this chapter has been auditing: a vivid, high-salience narrative (catastrophic superintelligent risk) is alleged to be drawing attention and resources away from a less dramatic but more directly

actionable structural problem (legal-accountability erosion), in a manner the four fidelity components could in principle assess — is the attention-capturing narrative's severity calibrated against its actual probability, does it depend on contingent and currently unrealized technical developments rather than direct, present mechanisms, and how much of its force depends on the prominence of the people discussing it rather than on demonstrated likelihood?

This book takes no position on how that audit would come out, and the case is included specifically to mark where the machinery's confident application from the previous case stops being available. Unlike the moral panic example, whether AGI risk is itself overstated, understated, or roughly proportionate to the attention it receives is a genuinely contested empirical and technical question among serious, informed people, not a case where one side's low fidelity is already established by available evidence. What can be said without resolving that question is narrower and still useful: the deflection structure being alleged — one salient risk narrative crowding out attention to a less salient but more tractable one — is a real and recurring pattern in how institutions and publics allocate limited attention, independent of whether superintelligence risk specifically is the salient narrative doing the crowding-out in this instance. The fidelity audit is the right tool for evaluating that structural claim case by case; it does not, by itself, settle the case in advance, and a book whose central caution is against treating borrowed authority as a substitute for demonstrated mechanism should not exempt its own use of a single contested example from that same caution.

This is offered as the book's most exportable single idea: a reader who rejects everything else argued here — the recursive compression principle, the weak-realist resolution of Chapter 17, even the relevance of intelligence as the book's organizing example — could still take the four-component fidelity audit and use it on a piece of rhetoric encountered tomorrow. Good reasoning, in any domain, consists substantially in learning which eliminations deserve authority. Bad reasoning consists, with notable regularity, in mistaking a constraint that is low on directness, low on calibration, low on closure, and high on borrowed authority for one that is none of those things.

17

Returning to Discovery vs. Constitution

In plain terms: Is intelligence discovered or constituted? This chapter closes that question — and it was actually closed quietly back in Chapter 5, not by any of the case studies in between. Some eliminations are forced by reality; some are pure rhetoric. Once you can tell them apart, intelligence can be neither purely discovered nor purely constituted — it's both, in different proportions for different parts of the concept.

17.1 The argument is already complete

Chapter 2 opened with a fork left deliberately unresolved: is intelligence discovered, a fixed target that better theories progressively approximate, or constituted, nothing more than whatever the elimination process happens to converge on, with no further fact about correctness available? The fork was left open because, at that point in the book, no machinery existed that could answer it honestly. That machinery now exists, and the fork can be closed — but it is worth being precise about exactly what closes it, because the temptation at this point in a book of this kind is to treat the accumulated weight of everything written since Chapter 2 as the argument, when in fact the argument was already complete by the end of Part II, before any of the case studies were run.

The reasoning is this. Some eliminations are demonstrably world-forced: a mathematical counterexample, a physical hazard that reliably produces the harm it threatens, a system's measurable incapacity to perform a task. These score high on every component of the Chapter 5 fidelity functional simultaneously, independent of who asserts them. Other eliminations are demonstrably rhetorical: amplified, low-directness, low-closure, authority-dependent — the empty bowl, the moral panic. Distinguishing the two requires the fidelity machinery; treating them identically, as either pure discovery (everything is world-forced, nothing is rhetorical) or pure constitution (everything is rhetorical, nothing is world-forced) would assert what the fidelity components were built specifically to deny. Therefore intelligence — or any concept subject to this kind of perspectival elimination — can be neither entirely discovered nor entirely constituted. Some of what survives elimination survives because the world will not permit otherwise. Some of what survives does so only because no

one has yet supplied, or no culture has yet been willing to supply, the constraint that would eliminate it.

Parts III through V of this book do not establish this argument. They stress-test it: Part III asks whether the underlying elimination dynamics behave sensibly in messy, real cases (Chapter 6, Chapter 7); Part IV asks whether a real cultural pedagogy already organizes itself around something like the fidelity distinction this book is formalizing (Chapter 9 through Chapter 13); Part V asks whether the recursive compression conjecture generalizes beyond intelligence to a second competence word (Chapter 14), and what would explain competence-word persistence if the conjecture is right (Chapter 15). None of these tests, however they turn out, can overturn the Chapter 2–5 argument, because that argument does not depend on any of them succeeding — and Part V is, in fact, the clearest available demonstration of this independence, since it did not turn out cleanly. Chapter 14 found only partial transport: the recognition/implementation shape recurs in creativity research, but creativity needed additional axes (domain expertise, social uptake) the intelligence case never surfaced, complicating rather than confirming the conjecture's specific two-axis form. Chapter 15 left its central conjecture explicitly unfinished, with no working cost model. Neither outcome touches the weak-realist resolution stated in this chapter. What narrows, given Part V's actual result, is only the generality claim of Chapter 13's conjecture — which is now better stated as "recursive compression is not unique to intelligence" than as "recursive compression takes the same shape everywhere," a real but more modest finding than a clean second success would have produced.

17.2 Stating the position with its actual content

Definition 17.1 (Weak-Realist Decomposition)**.** Restate Chapter 2's hard/soft split without the appearance of a formal limit it has not earned:

$$\Theta = H \cup S, \qquad H = \left\{ \theta \in \Theta \ : \ \begin{array}{l} \theta \text{ survives arbitrarily extended} \\ \text{sequences of high-fidelity elimination} \end{array} \right\}.$$

This is deliberately less formal-looking than an earlier version of this definition, which expressed H as a limit, $\lim_{t\to\infty}$, of a set-valued process. That phrasing was set aside because a limit of a set-valued sequence requires specifying what converges and in what topology — apparatus this book has not built and does not need. The set-builder form above says exactly what is meant and no more: H is whatever, as a matter of fact, keeps surviving no matter how long and how varied a sequence of high-fidelity perspectival contact is run against it. S is everything else — the dimensions that have moved, or could plausibly move, under different cultural or historical contact. Stating it this way is a small change, but it matters for the same reason the rest of this book has insisted on small changes of this kind throughout: the notation should never look more proven than the argument behind it.

Conjecture 17.2 (Persistence, demoted from "Theorem")**.** Elements of H remain invariant

under what might be called *cultural reparameterization* — a transformation of the local soft constraints S that a culture or era applies, while H is held fixed.

This cannot honestly be stated as a theorem, because the reparameterization operation it claims invariance under has not been formally defined anywhere in this book. Doing so would require specifying precisely what counts as a transformation of a culture's constraint set, and on what grounds H should be expected to be a fixed point of that transformation rather than merely a slow-moving one. Left as a conjecture, it restates the hard-core intuition in slightly more technical language without proving anything new about it — which is the honest status to assign it given what has actually been derived.

17.3 What this resolves and what it does not

The hard core of any competence bundle — the dimensions forced by contact with real, resistant failure — looks discovered, stable across cultures and eras because it is not up to cultures and eras to revise it. The soft periphery — caricature, rhetorical amplification, locally constituted emphasis — looks constituted, and shifts, because nothing forces it to stop. "Intelligence" names the union of both, and this is exactly why arguments about its definition have never fully resolved across a century of serious attempts: part of the disagreement is real, unsettled empirical territory about where the boundary of H actually sits, and part of it is a contest, often conducted by people who believe themselves to be doing the former, over which locally-constituted constraints in S get to count, rhetorically, as though they belonged to H. The discovery cycle that opened Chapter 1 is, on this account, not a series of failures to find the right definition. It is the visible record of that contest being fought out, repeatedly, across a century in which the tools to tell the two sides of the fight apart did not yet exist in stated form.

17.4 A Positive Criterion: Convergent Corroboration

In plain terms: So far this book only knows how to rule things out. But two completely independent fields sometimes arrive at the same structural claim with no contact between them — that's active confirmation, not mere survival, and deserves its own name. The worked case: a 1930s linguistic theory, confirmed decades later by neuroscience that had no idea the theory existed.

17.4.1 A case where theory arrived before the evidence that confirmed it

In the mid-1930s, Roman Jakobson and Nikolai Trubetzkoy proposed that phonemes — the basic sound units linguists had long treated as atomic — are not atomic at all. Each phoneme, on their account, decomposes into a small bundle of abstract binary distinctions: voicing, place of articulation, and several others. This was a purely theoretical claim, argued

on structuralist linguistic grounds, with no access to and no stake in any future neuroscience. There was, at the time, no method available that could have looked inside a living brain and checked.

Roughly seventy years later, intracranial electrode recordings taken from patients undergoing pre-surgical mapping for epilepsy — a technique unrelated in origin or motivation to 1930s phonology — found something specific: electrodes in auditory cortex do not specialize for individual phonemes. They specialize for exactly the kind of abstract distinctive features Jakobson and Trubetzkoy had proposed on paper, decades before anyone could test it. As Yosef Grodzinsky describes the finding in a 2026 lecture discussing the case, the coincidence is striking enough to call "sci-fi": a theoretical construct from one discipline, developed with no neuroscientific input, turned out to be the literal organizing scheme of a physical organ, confirmed by a second discipline that had no reason to vindicate the first.

This deserves to be stated narrowly and precisely, because it is tempting — and would be a fidelity violation in exactly the sense Chapter 5 warns against — to inflate the claim beyond what the evidence supports. What is established here is the auditory-cortex electrode finding itself: a specific structural prediction, made on theoretical grounds, confirmed by a specific later study using an independent methodology. It is not established, by this case alone, that distinctive feature theory has been continuously validated across phonology, acquisition, speech-error research, and aphasiology for a century — that broader claim may well be true, but this book has not verified it, and citing it as already demonstrated here would be borrowing exactly the kind of unearned narrative force this book's own machinery exists to catch.

What this case offers the book is not a confirmed grand history. It is a single, clean, well-documented instance of something the existing formalism has no name for.

17.4.2 Survival is not corroboration

Everything built in Chapters 3 through 5 is negative machinery. A constraint C_i rules candidates *out*; a candidate's standing in H is, so far, defined entirely by what it has not yet been eliminated by. Call this *survival*: $\mu_t(\theta)$ stays bounded away from 0 because nothing has yet driven it down.

The distinctive-features case is not a survival story. Nobody eliminated a rival candidate and left distinctive feature theory standing by default. Something stronger happened: a second, independent methodology, with no access to and no stake in the first theory, arrived at the same structural claim on its own. Call this *corroboration*: $\mu_t(\theta)$ rises, actively, because independent perspectives keep recovering it rather than merely failing to rule it out.

These are different processes, and the difference matters for what kind of confidence each one licenses. A candidate that has merely survived could still be wrong in a way no one has yet found the right perspective to detect — its standing is a statement about the limits of the search so far, not about the candidate itself. A candidate that keeps being independently recovered by methodologies that could not have influenced one another is in a different

epistemic position: the explanation that it is simply wrong has to also explain why unrelated investigative routes, with no contact between them, kept arriving at it anyway. That is a much harder coincidence to construct.

This motivates the section's central definition, stated before any of the supporting machinery beneath it:

Definition 17.3 (Convergent Corroboration)**.** A candidate θ is *convergently corroborated* when it remains recoverable across observational regimes — independently re-derived by perspectives whose mutual independence is high, especially when the regimes are separated by enough methodological or temporal distance that the later one could not have been shaped, even unconsciously, by the earlier one.

The Jakobson–Trubetzkoy case is the concrete instance motivating this definition, and it is worth being precise about which fact is doing the work. It is not merely that structural phonology and intracranial electrophysiology both, as it happens, endorse the same distinction. It is that the two methodologies are about as different as two ways of investigating language could be — and that the gap between them is large enough, roughly seventy years, that the second methodology did not yet exist when the first theory was proposed. Distinctive feature theory could not have been reverse-engineered from a foreknowledge of how electrode recordings would later turn out, because no one running the 1930s analysis had any such recordings to consult, and no one running the later study needed the 1930s theory to design their experiment around. The independence is not merely high; in this specific respect it is closer to structurally guaranteed.

17.4.3 Why independence is the construct the section actually depends on

The Lagrangian apparatus built later in this section is, on reflection, not where this section's real claim lives. The claim lives entirely in how independence is defined and measured, and it is worth stating plainly that the whole argument for convergent corroboration as a meaningful category stands or falls on this one construct.

Definition 17.4 (Independence)**.** For two perspectives P_i, P_j, let

$$\mathrm{Ind}(i, j) \in [0, 1]$$

denote one minus the fraction of shared evidentiary or methodological lineage between them, estimated from documented history: shared authors, shared instruments, shared funding sources, shared theoretical ancestry, or direct citation contact. $\mathrm{Ind}(i, j)$ is maximal when the two perspectives developed with no contact and no shared assumptions.

Two further, more specific conditions push $\mathrm{Ind}(i, j)$ toward its maximum, and both hold in the case motivating this section. *Methodological distinctness*: the two perspectives investigate the candidate through evidence of different kinds entirely — structural analysis of

phonological patterning versus electrical activity recorded directly from cortical tissue, sharing no instruments, no data, and no inferential machinery. *Temporal impossibility of contamination*: when the confirming methodology did not yet exist at the time the original claim was made, the confirming study cannot have been designed, even informally, around producing the expected result — there is no mechanism by which the earlier theory could have biased the later measurement, because the later measurement's entire technical apparatus postdates the theory it ends up confirming. This second condition is doing more work than it might first appear: it is close to a structural analogue of a pre-registered prediction, except stronger, since pre-registration only rules out a researcher consciously shaping a study around a known hypothesis, while temporal impossibility rules out the possibility of any such shaping occurring at all, conscious or not.

Strong versus weak independence

The discussion above gives two named conditions that push $\mathrm{Ind}(i, j)$ toward its maximum, but a reader trying to apply this machinery to a new case needs more than two extreme examples to work from. Is developmental psychology independent of linguistics? Is a second neuroscience laboratory, using the same scanner technology as the first, independent of it? These are exactly the questions a working taxonomy needs to answer, and a three-tier version, built from the two factors already in play — shared theoretical ancestry and shared instrumentation — is enough to organize most cases that will actually arise.

Weak independence: shared theory and shared instruments. Two laboratories using the same scanner technology, trained in the same theoretical tradition, and citing each other's prior work sit here. Agreement between them is closer to consensus than corroboration in the sense of Section 17.4's later subsection on that distinction — informative, but only weakly so, because a shared assumption or a shared instrumental artifact could produce the agreement independently of whether the underlying candidate is real.

Moderate independence: shared theory, different instruments. Two research programs that accept the same broad theoretical framework but reach their conclusions through different measurement techniques — behavioral testing versus neuroimaging, for instance, within a shared cognitive-science paradigm — rule out instrumental artifact as a common cause of agreement, but not theoretical bias: both could still be finding what their shared framework predisposed them to look for.

Strong independence: different theory, different instruments. This is the tier the Jakobson–Trubetzkoy case occupies, and it is the tier that does the most epistemic work, because it rules out both common causes of false agreement at once — neither a shared instrument nor a shared theoretical expectation can explain why two frameworks with nothing in common arrived at the same structure.

This taxonomy is offered as a practical aid to applying $\mathrm{Ind}(i, j)$ rather than as a further formal refinement of it — the three tiers are not claimed to exhaust the space or to be sharply bounded from one another, and real cases will often sit ambiguously between them.

What the taxonomy is meant to prevent is the more common error of treating any agreement between two named, separately-credentialed fields as automatically strong independence, when the relevant question is always the finer-grained one: how much theory and instrumentation do the two perspectives in question actually share, not merely what their fields are conventionally called.

Temporal independence

One further factor is worth separating out from the theory/instrument taxonomy above, because it behaves differently from either axis. The temporal impossibility of contamination condition already given is binary: either the confirming methodology existed at the time of the original claim, making some form of shaping conceivable, or it did not, ruling shaping out entirely. But time separation does additional work beyond this binary gate, and that additional work is graded rather than binary.

A recovery event fifty or seventy years after the original proposal is stronger corroborating evidence than a recovery event one week later, even holding theoretical and instrumental independence fixed, because longer separation makes alternative explanations for the agreement progressively less plausible: less chance of an unrecorded personal connection between the researchers involved, less chance that the original claim was still active enough in a field's working memory to bias a later study's design or interpretation, more opportunity for the original claim to have been quietly forgotten or fallen out of use, only to be rediscovered rather than confirmed. The Jakobson–Trubetzkoy case is, again, closer to the strong end of this than a typical example: roughly seventy years separate the original proposal from the confirming electrode recordings, a gap long enough that the original theoretical claim was not part of any live research conversation the electrophysiologists were responding to. Call this *temporal independence*, a graded factor that strengthens $\mathrm{Ind}(i, j)$ continuously as the gap widens, distinct from and additional to the binary impossibility-of-contamination condition: a recovery can satisfy the binary condition (the methodology genuinely did not exist yet) while still being temporally weak in this graded sense (the gap was short enough that the claim was still circulating actively when the confirmation arrived), and the strongest cases — this book's worked example among them — satisfy both.

Without a properly demanding definition of independence, convergent corroboration collapses into nothing more than consensus — two perspectives agreeing because they share assumptions, instruments, or community, which is a much weaker and more familiar phenomenon, fully susceptible to the same borrowed-authority failure mode Chapter 16 describes (an entire field can be wrong together, in agreement, for reasons having nothing to do with the world). With a properly demanding definition, the same surface pattern — multiple perspectives endorsing the same candidate — becomes something categorically different: rediscovery.

17.4.4 Recovery, alongside elimination

The existing machinery already has a quantity tracking how strongly perspectives argue *against* a candidate at a given perspective: $e_t(\theta)$, aggregated across perspectives as elimination pressure in the instability potential below. Convergent corroboration needs the positive counterpart.

Definition 17.5 (Endorsement)**.** For a perspective P_i and a candidate $\theta \in \Theta$, let

$$\varepsilon_i(\theta) \in [0, 1]$$

denote the degree to which P_i's own framework, evaluated on its own terms and stated prior to any corroborating evidence, specifically predicts θ — as opposed to merely permitting θ among many candidates the framework happens not to forbid.

Distinctive feature theory scores high on this measure with respect to the auditory-cortex finding: Jakobson and Trubetzkoy's framework did not merely fail to forbid feature-based neural organization. It specifically predicted that any physical encoding of phonemic contrast, wherever it was eventually found, would be organized by abstract feature rather than by raw acoustic identity — a falsifiable structural commitment made decades before the technology to check it existed.

Definition 17.6 (Recovery Count)**.**

$$R(\theta, t) = \left|\{(i, j) : i < j \leq n(t),\ \varepsilon_i(\theta) > \tau,\ \varepsilon_j(\theta) > \tau,\ \mathrm{Ind}(i, j) > \tau'\}\right|$$

for fixed thresholds $\tau, \tau' \in (0, 1)$ — the number of sufficiently independent perspective pairs that each, on their own terms, specifically endorse θ.

$R(\theta, t)$ is the direct positive counterpart to an elimination count: where elimination tallies how many perspectives a candidate has survived contact with, recovery tallies how many sufficiently independent perspectives have actively required it. A candidate with $R(\theta, t) = 0$ may still be entirely defensible — most candidates this book treats as part of H have never been tested for corroboration at all, only for survival, and the distinction is not meant to demote them. What $R(\theta, t) > 0$ adds is a second, stronger kind of standing that survival alone cannot confer.

With endorsement and independence both in hand, the weighted version follows:

Definition 17.7 (Pairwise and Aggregate Corroboration)**.**

$$K_{ij}(\theta) = \varepsilon_i(\theta)\, \varepsilon_j(\theta)\, \mathrm{Ind}(i, j), \qquad K(\theta, t) = \sum_{i<j\leq n(t)} K_{ij}(\theta).$$

This is genuinely computable, not merely evocative, and it is worth being explicit about what computing it would actually involve rather than leaving the claim abstract: take the

phoneme clustering implied by distinctive feature theory and the phoneme clustering implied by the electrode-response data, and measure their agreement above chance — an adjusted Rand index or normalized mutual information between the two independently-derived partitions would serve. That number is what $K(\theta)$ is measuring in this specific case. It has not been computed here; the underlying electrode data was not available to this book. The procedure is real and stated precisely enough to be carried out by someone with access to the data. The number itself remains an open empirical question, and the chapter should not be read as implying otherwise.

17.4.5 A proposed dynamical model

The definitions above — survival versus corroboration, independence, recovery count, and weighted corroboration — are where this section's actual content lives, and the Jakobson–Trubetzkoy case motivates and supports every one of them directly. What follows is different in kind and explicitly subordinate to it: a heuristic dynamical model, offered as a formal description of a process whose existence has already been established by the case above, not as the source of the section's authority. A reader who finds the Lagrangian unconvincing or unnecessary loses nothing of the section's substantive claim, which is fully stated already.

The operational definitions above are the part of this section the book can fully stand behind. What follows is different in kind: a heuristic dynamical model, built so that it reduces, in a stated limit, to machinery the book has already established. It earns inclusion only to the extent that reduction holds, and the reduction is offered here as a conjecture, not a completed derivation.

Treat $\mu_t(\theta)$ as a continuum field over Θ, using the conceptual-distance metric d_Θ from Chapter 1 to give nearby candidates meaning. Define the *instability potential*

$$V(\theta, t) = \underbrace{\sum_i w_i(t)\, e_i(\theta)}_{\text{elimination pressure}} - \lambda\, K(\theta, t), \qquad \lambda > 0.$$

Elimination pressure raises V, pushing plausibility down. Corroboration lowers V, carving a stable well a candidate can settle into.

Definition 17.8 (Belief-Field Lagrangian, proposed dynamical model)**.**

$$\mathcal{L}[\mu, \dot{\mu}, \nabla_\theta \mu] = \frac{\kappa}{2}\dot{\mu}(\theta, t)^2 \; - \; \frac{\sigma}{2}\big(\nabla_\theta \mu(\theta, t)\big)^2 \; - \; V(\theta, t)\, \mu(\theta, t).$$

The middle term is the one genuinely new structural addition here, and it formalizes something the book has so far only used informally: the discovery cycle of Chapter 1 narrows toward *neighborhoods* of definitions, not isolated points, which presupposes exactly the kind of smoothness this term enforces — two candidates close in concep-

tual distance should not differ wildly in plausibility absent specific evidence distinguishing them.
The associated Euler–Lagrange equation is

$$\kappa\,\ddot{\mu}(\theta,t) - \sigma\,\nabla_\theta^2\mu(\theta,t) + V(\theta,t) = 0.$$

Conjecture 17.9 (Overdamped Reduction)**.** In the regime where belief revision is slow relative to inertial effects ($\kappa\ddot{\mu} \approx 0$), the dynamics above reduce to gradient flow,

$$\sigma\,\nabla_\theta^2\mu(\theta,t) \approx V(\theta,t),$$

and a discrete-time, renormalized version of this flow recovers something structurally close to the corrected soft-elimination update of Chapter 5: μ decreases where elimination pressure dominates ($V > 0$), increases or stabilizes where corroboration dominates ($V < 0$), and diffuses smoothly across nearby candidates rather than updating each θ in isolation.

This is stated as a conjecture because it is plausible rather than demonstrated: the actual discretization and renormalization needed to confirm that this continuum model reproduces Chapter 5's update exactly, rather than merely resembling it qualitatively, has not been carried out. The gap is left visible deliberately, in keeping with the rest of this book's practice, rather than smoothed over by the elegance of the notation surrounding it.

Definition 17.10 (Conjugate Momentum and Hamiltonian)**.** With $\pi(\theta,t) = \partial\mathcal{L}/\partial\dot{\mu} = \kappa\dot{\mu}(\theta,t)$, the Legendre transform gives

$$\mathcal{H}[\mu,\pi] = \int_\Theta d\theta \left[\frac{\pi(\theta,t)^2}{2\kappa} + \frac{\sigma}{2}\big(\nabla_\theta\mu(\theta,t)\big)^2 + V(\theta,t)\,\mu(\theta,t)\right].$$

This is mechanically correct given the Lagrangian above — a genuine Legendre transform, not a separately invented expression — but its physical interpretation should not be overstated. Nothing in this book provides independent evidence that belief revision has well-defined momentum or conserves $\mathcal{H}$ in any literal sense. The Hamiltonian is included because it follows directly from the proposed Lagrangian and because the gradient-flow limit connecting back to Chapter 5 is the part doing real work; the energy-conservation language that comes bundled with Hamiltonian mechanics is borrowed for formal convenience, not asserted as a discovered property of how concepts actually move.

17.4.6 Corroboration is not agreement

One further confusion is worth heading off explicitly, since it is the most natural way to misread everything built in this section: convergent corroboration is not consensus, and the two should not be allowed to blur together.

Consensus concerns how many people, or how many restatements, endorse a claim. Corroboration concerns how many *independent routes* recover it. Three researchers who each repeat the same textbook argument may increase social confidence in a claim, but contribute little new epistemic information, because all three endorsements trace back to a single common source — in the terms of this section, $\text{Ind}(i, j) \approx 0$ for any pair among them, and the resulting agreement, however widely repeated, adds almost nothing to $K(\theta, t)$. Three genuinely independent methodologies that each, on their own terms and without contact with one another, recover the same candidate are in a different epistemic position entirely, because $\text{Ind}(i, j) \approx 1$ for each pair, and the probability of all three converging on the same structure through entirely unrelated routes, if that structure were not real, is correspondingly lower.

This is precisely why convergent corroboration measures something popularity cannot. A claim repeated by thousands of people can have low corroboration in this book's sense if every repetition traces back to the same original source — the moral panic of Chapter 16 is, among other things, an instance of exactly this: wide agreement, near-zero independence. A claim supported by only a handful of perspectives can have high corroboration if those perspectives reached it by genuinely unrelated paths.

The case motivating this section illustrates the distinction precisely, and it is worth restating its scope narrowly rather than letting the point above invite overreach. What this book has established is convergence between two regimes specifically: 1930s structural phonology and 2010s intracranial electrophysiology, with $\text{Ind}(i, j)$ close to maximal between that pair for the documented reasons already given. It would be a natural next step to suggest that a broader convergence — across language acquisition research, speech pathology, and other adjacent fields, in addition to the two regimes already discussed — would make the corroboration case stronger still, and that may well be true. But this book has not verified that broader convergence, and asserting it here, however tempting given how well it would fit the argument, would repeat exactly the kind of unearned extension this section's own earlier subsection warned against. The two-perspective case, stated at the scope it has actually been demonstrated, is already a strong and sufficient illustration of what corroboration measures that consensus alone cannot: not the size of a coalition, but the independence of the paths by which a structure is recovered.

17.4.7 What this section establishes and what it leaves open

The substantive claim of this section is now fully stated before any physics notation appears: survival and corroboration are different processes, convergent corroboration is well-defined once independence is given a sufficiently demanding operational meaning, and the Jakobson–Trubetzkoy case is a real, correctly attributed instance of it — with $\text{Ind}(i, j)$ close

to maximal for documentable reasons (methodological distinctness, temporal impossibility of contamination), and $R(\theta, t) > 0$ for that case on any reasonable choice of threshold. None of that depends on the dynamical layer that follows. The Lagrangian, its overdamped reduction, and the Hamiltonian are offered explicitly as a secondary formal description of a phenomenon the section has already established by other means — a conjecture-level model, built to connect to existing machinery in a specific limit, with that connection not yet rigorously confirmed. Distinguishing these two layers matters more here than almost anywhere else in the book, because the temptation to let the elegance of a Lagrangian formulation lend unearned credibility to the operational claims beneath it is exactly the kind of borrowed authority Chapter 16 was built to diagnose. A reader who finds the dynamical model unpersuasive loses nothing of the section's actual argument; a reader who finds it persuasive should be clear that its persuasiveness is borrowed from the case and the definitions that precede it, not the other way around.

18

Intelligence as Self-Description

In plain terms: Intelligence is a bundle of separable skills — recognizing what's right, doing it, and sub-skills beneath that — compressed into one convenient word. The strongest evidence for this isn't any case study; it's the reading experience itself: four times this book proposed an idea, watched it fail, and rebuilt something better, and once it failed only partially and said so. The book didn't just describe that process. It did it.

18.1 Returning to where the book began

The closing image of this book is the same image that opened Part IV: a Kalabari father places a chair, or an empty bowl, beside something genuinely edible, and a child learns, cheaply and safely, what an inadmissible answer looks like, long before the stakes are real. The device works by collapsing a complicated boundary — one that, examined closely, depends on causal directness, severity, closure, and source independence all at once — into something a child experiences as a single, obvious, unmissable contrast.

This book's closing claim is that the word "intelligence" performs the same operation for a culture that the counterfoil performs for a child. Intelligence is not a single faculty waiting to be measured, the way a century of confident, cycling definitions assumed it must be. It is also not an arbitrary social label attached to nothing real, which would make the entire discovery cycle of Chapter 1 a pointless argument about nothing. It is a structured, partially correlated, partially world-forced bundle of competencies — recognition and implementation, and beneath implementation a further bundle of inhibition, emotional regulation, effort, delay, and social cost, each separable and only partially correlated with the others — compressed, by convenience and by the ordinary economy of language, into a single word. The rigor available to a science of intelligence does not consist in finally finding the latent variable that word was secretly pointing at all along. It consists in learning, case by case, claim by claim, which parts of the compression are load-bearing — forced by contact with something that will not move — and which parts are decoration, asserted with more confidence than their fidelity actually supports.

18.2 Which parts of this book belong to H, and which to S?

A book that has spent eighteen chapters insisting on the distinction between a hard core forced by contact with reality and a soft periphery fixed only by local convention owes its reader one last application of that distinction: to itself. The audit below is offered in that spirit, not as false modesty and not as a hedge against criticism, but because skipping it would be the one place in this book where the standard applied everywhere else quietly stopped applying.

Candidates for this book's own hard core — claims that, on the book's own account, should be expected to survive contact with perspectives this book did not consult, cultures it did not consider, and critics it has not yet met — include the Irreversible Elimination theorem of Chapter 4, which is a structural fact about monotonic set-shrinking processes in general and does not depend on anything specific to intelligence, the Kalabari material, or any choice this book made about how to apply it; the recognition/implementation distinction of Chapter 10, which is now independently grounded in psychological literature (moral disengagement, self-regulation research) this book did not generate and could not have shaped; and the basic intuition behind fidelity weighting — that eliminating evidence varies in the authority it deserves, and that authority is not the same quantity as truth — which seems likely to hold regardless of which four components turn out to be the right operationalization of it.

Candidates for this book's soft periphery — choices that could plausibly have gone otherwise, without damaging the substantive claims — include the specific notation chosen throughout (Θ, μ_t, the particular letters assigned to the four fidelity components), which is convenient but not forced; the specific worked examples (the Kalabari counterfoil material, Room for the River, the distinctive-features case), each of which illustrates a claim this book makes without being the only possible illustration of it, and each of which could in principle be replaced by a different case without the underlying formal claims changing; and the particular pedagogical choices about chapter ordering, which parts were given full chapters and which were compressed into subsections, and how much weight any single example was given relative to another — all of which reflect what this project happened to have access to and happened to find compelling, not what was uniquely available to find.

Some candidates resist easy sorting, and naming them is more honest than letting the audit appear tidier than it is. The recursive compression conjecture sits uncomfortably between the two categories: it has survived one serious attempt at cross-domain testing (Chapter 14's partial, not clean, transport to creativity), which is more than mere assertion but less than the kind of repeated, varied contact this book's own definition of H requires before a claim earns that status. The field-theoretic apparatus of Section 17.4 is, by its own explicit admission, mostly periphery dressed in load-bearing notation — offered as a conjecture-level heuristic precisely because this book does not yet know whether it belongs in H or not, and says so rather than guessing.

18.3 The argument the book has just made about itself

One more thing remains to be said, and it is not a rhetorical flourish appended after the real argument has finished — it is itself the strongest available evidence for everything claimed above, and it is available because of how this book happens to have been written rather than because it was planned as a literary device from the outset.

Chapter 1 opened with a set of candidate definitions of intelligence and the observation that history's confident attempts to fix one have cycled rather than converged. Chapter 4 then took the first formal model this book proposed — pure intersection — and proved, rigorously, that it could not work: a single mistaken elimination under that model is permanent and unrecoverable. The model was repaired, not patched, by the fidelity-weighted machinery of Chapter 5. Chapter 7 took the first substantive empirical hypothesis this book proposed about a real domain — a single global category graph accumulating anomaly until it splits — and watched it fail against the developmental-psychology literature in a precise and informative way: production and comprehension dissociate, and the global graph does not exist as a single object to begin with. The model was repaired by subsystem indexing, constrained by an explicit anti-vacuity principle so the repair could not become unfalsifiable. Chapter 11 took a clean, sequential curriculum hypothesis — recognition training handed off to implementation training — and found it did not survive the Kalabari material's own earliest examples, where implementation challenges appear interleaved from early childhood. The model was repaired into a weighted, interleaved exposure curriculum. Chapter 12 took the most tempting available conclusion — that recognition and implementation are simply independent — and found the literature would not support flat independence either, repairing the claim into dissociable-but-not-orthogonal, with implementation itself fracturing into a vector under the same scrutiny.

Four times, across this book, a naive model was proposed, met a real case, failed in a specific and namable way, and was repaired into something less elegant and more defensible. A fifth instance is worth including here precisely because it does not fit the pattern as neatly as the first four, and a book auditing its own honesty should not quietly omit the one case that complicates its closing image. Chapter 14 took the recursive compression conjecture — built once, on intelligence alone — and asked whether it transported to creativity. It did not fail outright, the way the Chapter 7 category-graph model did. It also did not cleanly succeed, the way a tidier book might have reported. It transported partially: the recognition/implementation shape recurred, but creativity needed axes intelligence's case never surfaced. That is not a fifth clean repair. It is something more useful for the argument being made in this closing chapter — a demonstration that this book's own elimination process does not always resolve into a satisfying revision, and reports the cases where it does not as honestly as the cases where it does.

This is not incidental to the book's argument. It is the argument, performed rather than merely described. The reader's own working model of intelligence, across these eighteen

chapters, has moved through a sequence

$$\Theta_0 \to \Theta_1 \to \Theta_2 \to \cdots$$

structurally identical to the elimination dynamics Chapter 5 formalized — proposed, tested against a perspective, revised by something functionally equivalent to fidelity-weighted elimination, and carried forward into the next encounter. The method of this book is not merely a description of perspectival convergence offered from outside the process, the way an observer might describe a chemical reaction without being part of it. It is an instance of the process it describes, and the reader who has followed the four clean repairs and the one honestly incomplete one has just completed one full run of the very mechanism this book claims produces concepts like intelligence in the first place — partial results and all.

Proposition 18.1 (Reflexive Convergence)**.** *The sequence of models this book produces — at Chapters 4, 7, 11, 12, and 14 specifically, and in the book's overall trajectory from Chapter 1's open question to this closing chapter — instantiates the same naive-model, counterexample, fidelity-weighted-revision, repaired-model structure formalized in Chapters 3 through 5, including the honest case (Chapter 14) where the revision was partial rather than complete.*

This is stated as a proposition rather than a theorem because its support is direct demonstration — pointing at the specific instances where it happened, available for the reader to check against the chapters just cited — rather than an independent derivation from stated assumptions. That is the correct epistemic status for it. The claim that the book is congruent with a model of its own formation is an observation about the book's actual structure, verifiable by checking it, not a further mathematical result requiring proof on top of everything already proven. It is offered as the strongest evidence available for the book's central thesis, not because demonstration is logically stronger than argument, but because a book whose subject is the difference between earned and borrowed authority would have undercut its own central claim if its final chapter had tried to assert, rather than show, that the claim was true.

18.4 What this book is not claiming

Before closing with the question of what would count against this book's central claims, it is worth being equally explicit about what those claims are not, since a number of the strongest available objections to this book are objections to positions it does not actually hold — and a reader who came away believing otherwise would have been let down by an absence of clarity this section exists to close.

This book is not claiming that intelligence is unreal. The entire weak-realist argument of Chapter 17 depends on intelligence having a hard core forced by contact with reality; denying that intelligence is real, in any sense, would contradict the book's own central resolution of the discovery/constitution fork, not extend it.

This book is not claiming that intelligence, or any competence word, is infinitely decomposable. The recursive compression conjecture of Chapter 13 has been tested exactly twice — once on intelligence, once, partially, on creativity — and nothing in either result implies the decomposition continues without limit. Whether there is a principled stopping point, and what would mark it, is left as an open question this book does not address.

This book is not claiming that all perspectives are equally valid. The entire purpose of the fidelity functional in Chapter 5 is to deny exactly this — perspectives vary, often dramatically, in how much authority their constraints deserve, and treating them as interchangeable was the naive model's failure, not this book's position.

This book is not claiming that convergence guarantees truth. Chapter 5's Misweighted Elimination conjecture states a hoped-for average-case relationship between fidelity-weighted revision and error reduction, explicitly left unproven; nothing in this book asserts that a candidate surviving extended high-fidelity scrutiny is thereby guaranteed correct, only that it has earned a specific, articulable kind of standing that a merely-asserted candidate has not.

This book is not claiming that recursive compression applies to every concept, or that every single-word label conceals a recoverable bundle of sub-competencies in the way intelligence and creativity were shown to. The conjecture is stated narrowly, tested on exactly two cases, and Chapter 14 found the second case to transport only partially — a result this book has been explicit about not over-extending.

18.5 What would change my mind

Every theory examined critically in this book — pure intersection, the global category graph, the sequential curriculum, flat independence, the GPT-2/brain-correlation work of Chapter 16 — was asked, at some point, to state what observation would count against it. A book that has spent eighteen chapters applying that standard to other people's claims owes its reader one last application of it to its own, and this final section is that application, stated as plainly as the rest of the book has tried to state everything else.

The following observations, if made, would constitute genuine evidence against this book's central claims, not merely complications it could absorb the way it absorbed the failures of Chapters 4, 7, 11, and 12.

Recognition and implementation found inseparable. If careful, repeated investigation across many populations and many domains failed ever to produce a case of high I_R with low I_I, or the reverse — if the two quantities turned out, against the moral-disengagement and self-regulation evidence already cited, to move together with no genuine exceptions — the central formal move of Part IV would be unmotivated. The book's response would not be to redefine the terms until a gap reappeared; it would be to retract the dissociability claim and ask what, if anything, of the recognition/implementation split survives that retraction.

Recursive decomposition consistently failing. Chapter 14 found partial transport for one second case. If a wider, more careful survey of competence words — wisdom, charisma, willpower, and others — found that most resist any non-trivial decomposition into independently mea-

surable sub-components, the recursive compression conjecture would be in serious trouble, not merely narrowed. One partial success and several clean failures is a different evidential situation than the one this book currently reports, and would warrant abandoning the conjecture's general form rather than continuing to qualify it.

Corroboration adding no predictive value beyond elimination. The entire argument of Section 17.4 depends on convergent corroboration being a genuinely different and stronger relationship than mere survival. If it could be shown that candidates with high $R(\theta, t)$ are no more likely to remain stable under future high-fidelity scrutiny than candidates that have merely survived elimination without corroboration — that is, if corroboration turned out to predict nothing beyond what survival alone already predicts — the positive criterion this book introduced in Chapter 17 would have no work left to do, and the distinction between survival and corroboration, however conceptually clean, would not be empirically load-bearing.

Fidelity weighting increasing rather than decreasing error. The Misweighted Elimination conjecture of Chapter 5 predicts $\partial L/\partial w < 0$ for correctly calibrated constraints — that giving more authority to high-fidelity evidence should, on average, reduce error rather than increase it. This is, in principle, testable: if systems or reasoners that weighted evidence according to something like the four fidelity components consistently performed *worse* at tracking true conclusions than those that weighted all evidence equally, or randomly, the entire apparatus of Chapter 5 would have been solving a problem that does not exist, however intuitively compelling the four components seemed in isolation.

None of these four observations has been made. Stating them here is not a hedge against future criticism, and it should not be read as the book quietly preparing its own retreat. It is the last instance, in this book, of the same discipline applied to every model it has built and repaired: a claim that cannot say what would count against it has not yet earned the confidence with which it is held. This book has tried, throughout, to earn that confidence in pieces — two genuine theorems, a larger number of honest conjectures, and a closing willingness to say, in each case, exactly what kind of evidence the conjecture is still waiting on. The four observations above are what it is still waiting on now.

A

Formal Appendix: Load-Bearing Mathematical Objects by Chapter

In plain terms: This appendix is a reference table, not a new argument. It lists, chapter by chapter, every formal object the book relies on, labeled honestly according to what kind of claim it actually is: a *definition* (just naming something, no claim attached), a *proposition* (an easy, near-immediate consequence of the definitions), a *conjecture* (a precise, plausible, currently unproven claim), or a *theorem* (something actually derived from stated assumptions, with a complete proof). Applying this discipline honestly turns out to matter: most of the book's formal content is definitions and conjectures, with only two genuine theorems in all eighteen chapters. That is the correct and healthy shape for a book whose subject is, in part, the difference between earned and borrowed authority — a book that called everything a "Theorem" would be committing, in its own mathematics, exactly the error its own Chapter 5 exists to diagnose.

Applied honestly, most of the book's formal content turns out to be definitions and conjectures, with a small number of genuine propositions and exactly one piece of mathematics that earns the word "theorem" without qualification — joined by one more, a confirmation of corrected machinery in Chapter 5, that is modest but also genuine. That distribution is the correct one for a book whose subject is, in part, the difference between earned and borrowed authority.

Part I — The Problem of the Definition

Chapter 1

Definition (Definition Space). $\Theta = \{\theta_1, \theta_2, ...\}$, the space of candidate definitions of intelligence, equipped with a conceptual-distance metric $d_\Theta(\theta_i, \theta_j)$.

Conjecture (Definition-Space Expansion). In the historical record of a contested concept, $|\Theta_t|$ tends to increase before any convergence mechanism causes it to decrease. Motivates the need for elimination machinery in Part II; does not establish that machinery is necessary.

Chapter 2

Definition (Hard/Soft Decomposition). $\Theta_t = H_t \cup S_t$, where H_t is the hard core and S_t is the soft periphery.

Conjecture (Differential Stability). $dH_t/dt \to 0$ as evidence accumulates, while dS_t/dt may remain nonzero indefinitely. The book's central organizing intuition, not yet a derived result — it depends on what "evidence accumulation" formally does to H and S, which is not specified until Chapter 5.

Part II — The Formal Machinery

Chapter 3

Definition (Naive Elimination). $\Theta_n = \bigcap_{i=1}^{n} C_i$, where each C_i is a perspective-supplied constraint.

Proposition (Monotonicity). For pure intersection, $\Theta_{n+1} \subseteq \Theta_n$ for all n. Immediate from the definition of intersection.

Definition (Perspective, added to Chapter 3). A perspective P is any source — observer, culture, institution, developmental stage, research program, or system — capable of producing evidence bearing on Θ. Each perspective induces a constraint via $\Phi : \{\text{perspectives}\} \to \{\text{constraints on } \Theta\}$, with $C_i = \Phi(P_i)$. This closes a gap in an earlier draft: a book titled around *perspectival convergence* had constraints (C_i) but no formal object named "perspective" until this definition was added.

Chapter 4

Definition (Belief Measure). $\mu_t : \Theta \to [0, \infty)$, a plausibility assignment over candidate definitions, updated by $\mu_{t+1} = U(\mu_t, C_t)$ for some update rule U, rather than by set intersection.

Theorem (Irreversible Elimination). If a sequence of updates is strictly monotonic and shrinking — $\Theta_{t+1} \subseteq \Theta_t$ for all t — then mistaken elimination of the true candidate θ^* can never be repaired by any subsequent update in the sequence. *The strongest piece of mathematics in the book*: a real theorem, with explicit assumptions, a complete and non-trivial proof, and a conclusion that does actual argumentative work.

Chapter 5

Definition (Fidelity Functional). $w_i = f(d_i, s_i, c_i, q_i)$, where d_i is causal directness, s_i is severity calibration, c_i is closure, and q_i is source independence. Normalized form: $w_i = d_i \cdot s_i \cdot c_i \cdot q_i$, with each factor scaled to $[0, 1]$.

Definition (Soft Elimination, corrected). Originally proposed as $\mu_{t+1}(\theta) = \mu_t(\theta)(1 - w_t e_t(\theta))$, which does not preserve total probability mass under repeated application. Corrected via renormalization: $\mu_{t+1}(\theta) = \mu_t(\theta)(1 - w_t e_t(\theta))/Z_t$, with $Z_t = \sum_\theta \mu_t(\theta)(1 - w_t e_t(\theta))$.

Theorem (Fidelity Stability, under the corrected update). If $0 \leq w_t e_t(\theta) \leq 1$ for all θ and $Z_t > 0$, then $\mu_t(\theta) \geq 0$ and $\sum_\theta \mu_t(\theta) = 1$ for all t. A real, if modest, theorem — it confirms the corrected update rule behaves well, which the original uncorrected version did not.

Conjecture (Misweighted Elimination). $L = \mathbb{E}[(\hat{\mu} - \mu^*)^2]$; for correctly calibrated constraints, $\partial L / \partial w < 0$. The formal target the rest of the book gestures toward whenever it invokes "low-fidelity counterfoils mistaken for hard eliminations," but not yet derived: requires a stated model relating $e_t(\theta)$ to the true μ^*, which the book has not specified.

Part III — How Convergence Actually Behaves

Chapter 6

Definition (Interface Lag). For coupled subsystems S_i, S_j with closure times t_i, t_j, $\Delta t_{ij} = |t_i - t_j|$.

Conjecture (Coupling–Lag Relation). $\Delta t_{ij} \propto C_{ij}^{-1}$, where C_{ij} measures coupling strength. Checked only qualitatively against one institutional case; no quantitative estimate of C_{ij} attempted.

Definition (Nominal vs. Operational Closure). $N(t) \in \{0, 1\}$, declared closure; $O(t) \in \{0, 1\}$, actual operational closure. A *closure illusion* is a state where $N(t) = 1$ while $O(t) = 0$.

Chapter 7

Definition (Subsystem-Indexed Repair). Replace a single $R_n(X)$ with $R_n^{(k)}(X)$ for subsystem k.

Principle (Anti-Vacuity). A subsystem split into index k is admissible only if there exists an independent measurement method M_k for that subsystem, established by existing methodology rather than introduced solely to rescue a failed prediction. A methodological commitment, not a mathematical claim — it has no proof because it isn't the kind of statement that has one.

Chapter 8

Definition (Cross-Domain Structure). $\mathcal{S} = (\Delta t, N/O)$, the joint pattern of interface lag and closure status observed in a domain.

Observation (Structural Resemblance, demoted from "Cross-Domain Invariance"). The same formal structure $\mathcal{S}$ was observed in two hand-selected domains. The originally proposed inequality $P(\mathcal{S} \mid D_1, D_2) > P(\mathcal{S} \mid D_1)P(\mathcal{S} \mid D_2)$ presupposes a sampling process that does not exist here; with $n = 2$ non-randomly-chosen cases, no probability statement of this form is estimable.

Part IV — The Pedagogy of Boundaries

Chapter 9

Definition (Counterfoil Pair). A pair (a, c) with a admissible, c inadmissible, and distance $d(a, c)$ deliberately maximized subject to pedagogical relevance.

Analogy (demoted from "Margin Theorem"). Counterfoil pedagogy resembles high-margin training contrasts in statistical classification. An evocative structural parallel, not a derived result — the formal guarantees behind real margin-based learning theory have no established analogue for human concept acquisition via cultural pedagogy.

Chapter 10

Definition (Recognition and Implementation Competence). $I_R = P(f(x) = y)$; $I_I = P(\pi(x) = a \mid f(x) = a)$. The conditional definition cleanly separates implementation failure from recognition failure by construction.

Proposition (Dissociation). High I_R does not imply high I_I. Supported constructively, via the moral disengagement literature, not by formal proof.

Chapter 11

Definition (Curriculum Trajectory). A stage-indexed map $s \mapsto (d_p(s), m(s))$.

Conjecture (Interleaved Exposure, relabeled). Effective exposure at stage s is weighted: $E_s = \alpha_s I_R + \beta_s I_I$, with $\alpha_s > \beta_s$ early. Motivated by, but not established by, roughly a dozen illustrative, non-randomly-sampled cases.

Chapter 12

Definition (Transfer Matrix). $T = \begin{pmatrix} T_{RR} & T_{RI} \\ T_{IR} & T_{II} \end{pmatrix}$.

Proposition (Orthogonality Failure). If $T_{IR} > 0$ or $T_{RI} > 0$, then I_R and I_I are not orthogonal. True essentially by definition.

Proposition (Dissociability). If there exist population subsets with $I_R \gg I_I$ and others with $I_I \gg I_R$, then the two capacities are non-identical. Also near-definitional.

Chapter 13

Definition (Recoverable Compound, replacing "Recursive Compression Operator / Decompression Theorem"). A competence label ℓ is recoverable-as-compound if there exist measurable $(x_1, \ldots, x_n)$, $n > 1$, and an aggregation map $g : \mathbb{R}^n \to \mathbb{R}$ such that $\ell \approx g(x_1, \ldots, x_n)$, with the x_i not mutually redundant. Replaces an earlier, mathematically unsound formulation ($C : \mathbb{R}^n \to \mathbb{R}$ paired with a claimed inverse C^{-1}): a many-to-one map has no general inverse.

Conjecture (Recursive Compression). Whenever a culture names a competence with a single word, closer inspection tends to reveal that the label is recoverable-as-compound. Currently supported by exactly one worked case, pending Chapter 14.

Part V — A Second Pass

Chapter 14

Conjecture (Generalization Test). Creativity is recoverable-as-compound in the same sense. *Outcome: partially confirmed.* A recognition/implementation-like split recurs, but creativity carries additional axes (domain expertise, social uptake) the intelligence case did not surface.

Definition (Structural Isomorphism, as a test criterion). A map φ from I-components to C-components "preserves decomposition structure" if $\varphi(I_R)$ corresponds to a recognition-type component of creativity and $\varphi(I_I)$ to an implementation-type component. Offered as the chapter's evaluation criterion, not a proven isomorphism.

Chapter 15

Conjecture (Compression Efficiency, with explicit open cost model). Communication cost favors compression — $L(\text{profile}) \gg L(\text{single label})$ — and a label persists whenever savings in description length exceed the predictive error introduced: $\Delta L > \Delta E$. Gestures at a real formal tradition (rate–distortion theory, minimum description length) but lacks the cost model that would let the inequality be evaluated. Left explicitly as the chapter's open formalization target.

Part VI — Resolving the Fork

Chapter 16

Definition (Borrowed Authority). $B = 1 - q$, where q is the source-independence factor from Chapter 5's fidelity functional.

Definition (Counterfoil Distortion Index). $D = (\text{depicted severity})/(\text{actual severity})$.

Proposition (checked for circularity, relabeled from "Fidelity Theorem"). World-forced eliminations are characterized by high w; rhetorical eliminations by high $B = 1 - q$. Partly definitional, since q is one factor in w; the substantive claim is that rhetorically-amplified counterfoils tend to score low on q *and* d, s, c simultaneously — joint co-occurrence, not the definitional relationship, is the chapter's real claim.

Chapter 17

Definition (Weak-Realist Decomposition, restated without overclaiming formal apparatus). $\Theta = H \cup S$, where H is the set of candidates $\theta \in \Theta$ that survive arbitrarily extended sequences of high-fidelity elimination. An earlier draft expressed H as $\lim_{t\to\infty}$ of a set-valued process; set aside because such a limit requires specifying what converges and in what topology, which the book neither builds nor needs.

Conjecture (Persistence, relabeled from "Theorem"). Elements of H remain invariant under a (currently undefined) operation of "cultural reparameterization." Cannot be a theorem because the reparameterization operation has not been formally defined.

Definition (Survival vs. Corroboration; reordering principle for Section 17.4). Survival: $\mu_t(\theta)$ stays bounded away from 0 because nothing has eliminated it (purely negative — the standing this book's Chapters 3–5 machinery already supports). Corroboration: $\mu_t(\theta)$ rises because independent perspectives actively recover it. Distinguished explicitly because the original draft of this section did not separate them, risking collapse of corroboration into mere consensus.

Definition (Convergent Corroboration). A candidate θ is convergently corroborated when recoverable across observational regimes — independently re-derived by sufficiently independent perspectives, especially when separated by enough time or method that the later one could not have been shaped by the earlier one. Stated as the section's lead definition rather than left implicit until after the formalism.

Definition (Independence, with two named sub-conditions). $\mathrm{Ind}(i,j) \in [0,1]$, estimated from shared lineage history. Two conditions push it toward maximum: *methodological distinctness* and *temporal impossibility of contamination* (when the confirming methodology postdates the original claim, no shaping — conscious or otherwise — is possible; stronger than pre-registration). Flagged as the construct the entire section actually depends on: without a demanding independence criterion, corroboration collapses into consensus, fully vulnerable to the Chapter 16 borrowed-authority failure mode.

Definition (Endorsement, Recovery Count, Pairwise/Aggregate Corroboration; Section 17.4). $\varepsilon_i(\theta)$ measures whether a perspective specifically predicts θ rather than merely permitting it. $R(\theta,t)$ counts sufficiently independent perspective pairs that each specifically endorse θ above threshold — the direct positive counterpart to an elimination count, distinct from and prior to the weighted $K_{ij}(\theta) = \varepsilon_i(\theta)\varepsilon_j(\theta)\mathrm{Ind}(i,j)$, aggregated as $K(\theta,t)$. Introduced to represent positive confirmation across independent perspectives, a relationship the purely negative elimination machinery of Chapters 3–5 cannot express. Illustrated by the Jakobson–Trubetzkoy distinctive-features case (theoretical claim, 1930s) and its independent confirmation by intracranial electrophysiology (2010s), as discussed by Yosef Grodzinsky. Computable in principle (e.g., via adjusted Rand index between independently-derived clusterings); not computed in this book.

Definition (Belief-Field Lagrangian and Hamiltonian, proposed dynamical model — explicitly secondary). $\mathcal{L} = \frac{\kappa}{2}\dot{\mu}^2 - \frac{\sigma}{2}(\nabla_\theta \mu)^2 - V\mu$, with instability potential V combining

elimination pressure and corroboration; Legendre transform gives $\mathcal{H}$. This layer is explicitly framed as a formal description of a phenomenon the operational definitions above have already established, not as the source of the section's claim — a reader unpersuaded by the dynamics loses none of the section's substance.

Conjecture (Overdamped Reduction). The Euler–Lagrange dynamics, in the slow-revision limit, plausibly reduce to something structurally close to Chapter 5's corrected soft-elimination update. Not rigorously confirmed; the discretization/renormalization needed to establish exact correspondence has not been carried out.

Chapter 18

Proposition (Reflexive Convergence, relabeled from "Theorem"). The sequence of models the book itself produces, $\Theta_0 \to \Theta_1 \to \Theta_2 \to \cdots$, instantiates the same elimination-and-repair structure defined formally in Chapters 3–5, as demonstrated concretely at Chapters 4, 7, 11, 12, and 14 (the last partially). A proposition supported by direct demonstration — pointing at the specific instances where it happened — rather than a theorem with independent assumptions and proof.

Summary table

Chapter	Definitions	Propositions	Conjectures	Theorems
1	1	—	1	—
2	1	—	1	—
3	2	1	—	—
4	1	—	—	**1 (genuine)**
5	2	—	1	**1 (genuine)**
6	2	—	1	—
7	1	—	—	— (1 Principle)
8	1	—	—	— (1 Observation)
9	1	—	—	— (1 Analogy)
10	1	1	—	—
11	1	—	1	—
12	1	2	—	—
13	1	—	1	—
14	1	—	1	—
15	—	—	1	—
16	2	1	—	—
17	9	—	2	—
18	—	1	—	—

Two genuine theorems out of eighteen chapters, both load-bearing (Chapter 4 motivates the entire move to soft elimination; Chapter 5 confirms the corrected version of that ma-

chinery is well-behaved), surrounded by a much larger population of definitions and conjectures. That is the honest shape of the project at this stage, and it is a healthier one than a book with eighteen chapters each claiming a "Theorem" would have been — a reader can now see exactly where the book has actually proven something, where it has only defined terms, and where it is making a precise, falsifiable bet it has not yet cashed in.

References and Sources

In plain terms: This book draws on a handful of named, attributable sources (Wariboko's ethnography, the Jakobson–Trubetzkoy distinctive-features story as discussed by Grodzinsky) and a wider, unnamed body of literature consulted while developing the institutional and developmental case studies of Part III and the transfer literature of Chapter 12. This chapter lists what can actually be verified and cited properly, and is honest about the limits of that verification rather than presenting borrowed authority dressed as a complete scholarly apparatus — which would be exactly the kind of failure this book's own Chapter 5 was built to catch.

Directly cited and attributed in the text

Wariboko, Nimi. "Counterfoil Choices in the Kalabari Life Cycle." *African Studies Quarterly*, Volume 3, Issue 1 (1999). Center for African Studies, University of Florida. ISSN: 2152-2448. The source for the entire Kalabari case study developed in Chapters 9 through 13, including the life-cycle stages, the bibife and chieftaincy-installation examples, and Wariboko's own explicit framing of counterfoil choice as testing cultural "wisdom" rather than intelligence — a distinction this book respects throughout rather than overriding.

Grodzinsky, Yosef. "How Deeply Human is Language? Chomsky, the Brain, and the..." Lecture, 2026. The source for the distinctive-features case discussed in Section 17.4, including the auditory-cortex electrode finding, the Lunar Society framing referenced in earlier development of this project, and the Fodor–Hinton modularity anecdote. This is cited from a lecture transcript rather than a peer-reviewed publication; the transcript available to this book did not include complete venue or institutional metadata beyond what the lecture itself states, and this citation should be read accordingly — as attribution to a specific, identifiable lecture, not as a claim to the editorial standing of a published paper.

Jakobson, Roman, and Nikolai Trubetzkoy. Foundational work on distinctive feature theory, Prague School phonology, 1930s. The theoretical claim discussed in Section 17.4 — that phonemes decompose into abstract binary distinctive features — is associated with this collaboration and period in the history of linguistics, as discussed by Grodzinsky in the lecture cited above. This book has not traced the claim to a single original paper with confirmed publication details, and the attribution should be read as historically accurate at the level of the school and period (the Prague Linguistic Circle's development of distinctive feature theory through the 1930s, associated most closely with Trubetzkoy's subsequently published

Grundzüge der Phonologie, 1939, and Jakobson's later collaborative work) rather than as a precise bibliographic citation to one source.

"Is AI a Threat?" Panel discussion, the Royal Institute of Philosophy, 26 June 2026. Chaired by Edward Harcourt (Academic Director, Royal Institute of Philosophy, and Director of the Oxford Institute for Ethics in AI), with panellists Federica Lucivero (BRAID Fellow at the Ada Lovelace Institute and Associate Professor in Ethics of Technology at the Ethox Centre, University of Oxford), Jens Munch (technology and AI strategist, entrepreneur, and investor), Priyanka Suresh (Data Scientist in AI Safety, Google DeepMind), and Shannon Vallor (Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence, Edinburgh Futures Institute, and author of *The AI Mirror*). The source for the closure-illusion illustration of Chapter 6, the delegated-competence discussion of Chapter 10, and the deflection case of Chapter 16. Cited from a panel transcript rather than a peer-reviewed publication, with the same caveats given for the Grodzinsky lecture above.

Consulted in developing the case studies of Part III and Chapter 12

The following sources informed the institutional case study of Chapter 6, the developmental case study of Chapter 7, and the transfer literature discussed in Chapter 12. They are not individually name-checked in the prose of those chapters, which draw on them collectively rather than attributing specific claims to specific papers; they are listed here for transparency and so a reader can verify the underlying material independently. Author names are given where confirmed; where a source's byline could not be confirmed with confidence, only the title, venue, and year are listed rather than a guessed attribution.

Room for the River and Dutch flood governance

"Room for Rivers: Risk Reduction by Enhancing the Flood Conveyance Capacity of The Netherlands' Large Rivers." *Geosciences*, MDPI (2018).

"Room for the River: Innovation, or Tradition? The Case of the Noordwaard." Book chapter, 2020.

"Dutch Regional Water Authorities: Institutional Change and Continuity versus Environmental Challenges." Conference paper, European Society for Environmental History, Vrije Universiteit Amsterdam (2022).

"River Basin Approach and Integrated Water Management: Governance Pitfalls for the Dutch Space-Water-Adjustment Management Principle." Journal article (2006).

"Room for the River, No Room for Conflict." Research paper (2018).

"Remaking 'Nature': The Ecological Turn in Dutch Water Management." Research paper.

"Room for the River." *Wikipedia*, accessed 2026. Used for general chronology (the 1993/1995 floods, the 2006 Spatial Planning Key Decision, 2007–2015/2018 construction window) cross-checked against the sources above.

Child language development and category formation

Horst, J. S., Oakes, L. M., and Madole, K. L. "Time Course of Visual Attention in Infant Categorization of Cats versus Dogs: Evidence for a Head Bias as Revealed through Eye Tracking." *Child Development* 76, no. 3 (2005): 614–631.

"Overextensions in Comprehension and Production Revisited: Preferential-Looking in a Study of Dog, Cat, and Cow." *Journal of Child Language*, Cambridge University Press.

"A Computational Theory of Child Overextension." *Cognition*, Elsevier (2020).

Self-regulation, executive function, and moral disengagement

Nash, K., Schiller, B., Gianotti, L. R. R., Baumgartner, T., and Knoch, D. "Electrophysiological Indices of Response Inhibition in a Go/NoGo Task Predict Self-Control in a Social Context." *PLOS ONE* 8, no. 11 (2013): e79462.

"Self-Control and Cooperation in Childhood as Antecedents of Less Moral Disengagement in Adolescence." *Development and Psychopathology*, Cambridge University Press (2021).

"Annual Research Review: On the Relations among Self-Regulation, Self-Control, Executive Functioning, Effortful Control, Cognitive Control, Impulsivity, Risk-Taking, and Inhibition for Developmental Psychopathology." Review article.

"Moral Disengagement." *Wikipedia*, accessed 2026. Used for the general framing of Bandura's social-cognitive account of moral disengagement; the underlying theoretical framework originates with Albert Bandura's work on social cognitive theory and moral agency.

A note on what this list does not include

This book also draws, in Chapter 14, on the general shape of the existing creativity research literature — componential and systems-level accounts distinguishing divergent generation, evaluative judgment, domain expertise, persistence, and social uptake. That chapter is explicit that this reflects general, field-level consensus rather than any specific paper, and no individual source is cited there for the same reason none is listed here: attributing a field's general shape to one paper would itself be a fidelity violation of the kind this book has tried consistently to avoid. A reader wanting to pursue that literature further is better served by a current survey of componential creativity research than by a single citation this book is not in a position to select responsibly.