Contents

Defense, Self, and Speech: From Biological Vectors to Linguistic Form[1]

Paper II — Is Compensatory-Defensive Speech Computable? The Combinatorial Architecture of a Finite Operational Basis

Author: Maria Svet, Mindloom Cognitive Architecture Lab **Acknowledgment:** Structural development conducted in extended dialogue with Claude (Anthropic); see final Acknowledgments section. **Series:** *Defense, Self, and Speech: From Biological Vectors to Linguistic Form*, Paper II of VI **Working name between authors:** *Architecture of Protective Speech* by Masha and Claude **Version:** v0.8 (preprint) **Target venues:** PsyArXiv (English) · PsyArXiv / Russian preprint server (Russian) · expanded version to *Frontiers in Psychology* or *Cognitive Science*

0. How this paper sits in the series

Paper I established that the four canonical defensive responses — fight, flight, fawn, freeze — are preserved as directional vectors when organisms acquire language, and that a specific subclass of defensive speech, which we term *compensatory-defensive*, is structurally organized by the pre-emption of predicted shame against a rigid social-normative prior. Paper II takes up the methodological follow-up question: is this architecture formalizable in a way that makes compensatory-defensive speech, at least partially, computable? We argue that it is, through an integrated predictive-processing account: compensatory-defensive speech is generated from a finite alphabet of speech-level operations organized by the four biological vectors, whose temporal patterns realize four recurrent precision regimes within a hierarchical generative model. The resulting hypothesis space is both bounded and theoretically principled. This is the methodological foundation on which the remainder of the series rests: Paper III describes the engine that operationalizes this hypothesis space as an ontology-bounded classification architecture; Paper IV tracks trajectories within the

[1]Mindloom Cognitive Architecture Lab, 2026 info@mindloom.ch

regime-state space; Papers V and VI empirically validate and apply the approach on human and machine speech respectively.

1. Abstract

Paper I of this series argued that four canonical defensive responses — fight, flight, fawn, freeze — are preserved as directional vectors when organisms acquire language, and that a specific subclass of defensive speech (which we term *compensatory-defensive*) is organized by the pre-emption of predicted shame against a rigid social-normative prior. Within the predictive-processing framework, such defense is a form of hierarchical Bayesian inference operating on protected priors. The present paper asks the methodological follow-up: is this architecture formalizable in a way that makes compensatory-defensive speech *computable*?

We argue that it is, and the argument proceeds along two coupled axes.

First axis: a finite operational basis. Compensatory-defensive speech is generated from a bounded alphabet of **speech-level operations** — not templates, not fixed phrases — drawn from a constrained inventory organized by the four biological vectors, and composed according to context-sensitive but tractable rules. The sequences that result are in principle unlimited; the operations that produce them are not. This is the relation that holds between four nucleotide bases and the genomic library, or between the finite rules of chess and the astronomical space of possible games. We present four convergent lines of evidence: a *metabolic* argument (sustained compensatory defense on de novo assembly would be allostatically prohibitive, given the intersection of elevated base cost, absence of release, and continuous operation), a *clinical* argument (a century of independent observational traditions — psychoanalytic, object-relations, attachment, social-cognitive, trauma-clinical — have converged on a small number of recurring forms, a convergence documented here through four cross-tradition triangulations), a *predictive-computational* argument (priors are reused because prediction is an economy, and defensive priors are reused with particular rigidity because they are protected from updating), and a *developmental* argument (defenses crystallize into recurring forms through developmental sedimentation rather than being reinvented per situation).

Second axis: precision regimes as emergent attractors. The same operations, viewed at the level of temporal pattern and precision allocation, cluster into four recurrent architectures of active inference: **DEFENCE** (elevated prior precision on self-protective priors, suppressed likelihood precision on disconfirmatory channels, reduced epistemic weight), **COGNITIVE** (balanced, context-sensitive precision allocation supporting evidence-updating and cooperative repair), **ILLUSION** (biased prior means with selective reduction of likelihood precision, self-sustaining without acute threat), and **META** (higher-order inference over precision itself, enabling reflexive modulation of the lower levels). These four regimes are not categories imposed on the data; they are attractor-like configurations of precision allocation within a single hierarchical generative model. The operations of the first axis are the observable speech-level outputs; the regimes of the second are the dynamical patterns they trace. We present the conceptual structure of the four regimes in the main text and the formal characterization — precision hyperparameters, a linear-Gaussian toy model, and five falsifiable

predictions — in Appendix A.

Together, these two axes establish that the hypothesis space for compensatory-defensive speech is *both finite and theoretically principled*. This is the methodological result on which the remainder of the series depends. Paper III describes the engine that operationalizes this hypothesis space as a neuro-symbolic classification architecture with *structure-constrained inference* — rule-based layers defining a bounded hypothesis space within which LLM judgment is exercised. Papers V and VI empirically validate and apply the approach. What we offer here is not an engine, not empirics, and not applied methodology — it is the theoretical warrant for the classification work that follows: a demonstration that the space within which that classification operates is finite, structured, and derivable from independent theoretical considerations rather than being imposed by engineering convenience.

We close by stating explicitly what this paper does not claim. It does not claim that speech in general is combinatorial, or that all precision dynamics in discourse conform to four regimes. The claim is narrower and principled: a specific subclass of speech — compensatory-defensive, produced under predictive-shame dynamics — admits finite-basis formalization with a well-defined regime structure, and this formalization is what authorizes the downstream engineering.

Keywords: defensive speech, computability, combinatoriality, finite basis, predictive processing, active inference, precision regimes, allostatic load, clinical stereotypy, speech operations, annotation ontology, ontology-bounded classification.

2. Introduction

2.1 The question

The question this paper takes up is simple to state and unexpectedly demanding to answer. If compensatory-defensive speech is, as Paper I argued, a coherent phenomenon rooted in preserved biological vectors and organized around the pre-emption of predicted shame, can we *formalize* it? Can we specify its operations with enough precision that a system — human annotator, computational classifier, or trained clinician — could recognize, label, and reason about instances of it in a principled way?

The question matters because the answer determines what becomes possible downstream. If compensatory-defensive speech is formally intractable — if it is fundamentally creative, unbounded, irreducible — then everything Paper I proposed remains theoretical literature, interesting but inert. If it admits partial formalization, then an ontology is possible; if an ontology is possible, an engine is possible; if an engine is possible, empirical studies, clinical tools, and safety evaluations for language models all become reachable. The computability question is not an ornamental aside. It is the hinge on which the remainder of the series turns.

2.2 The predictive-processing framework

The question of computability cannot be addressed in the abstract. It has to be addressed within a framework that specifies what the speaking brain is doing when

it produces speech and what kind of thing, at base, speech is. For this purpose we adopt the predictive-processing framework (Friston, 2010; Clark, 2013; Hohwy, 2013), within which the brain is modeled as a hierarchical generative model that continuously predicts its sensory inputs, compares predictions to incoming signals, and updates through prediction-error minimization. Action, in this framework, is not a separate category from perception; both are forms of inference, with action being the execution of policies that minimize expected free energy (Friston et al., 2017; Parr et al., 2022). The framework is not ornamental; it is the operating theory within which the argument of this paper is made. Where the three framings we will consider in §2.3 are candidate *theories of how language works as a computable object*, predictive processing is the framework within which we specify what the speaking agent is and what its outputs are.

This framework has specific consequences for the computability question that matter for what follows. Within predictive processing, sustained behaviors tend toward cached implementation. This is a direct consequence of the free-energy principle: an organism that reconstructed its response on every occasion would incur prediction-error and metabolic costs that the architecture is explicitly designed to minimize. Cached priors — stable probabilistic expectations over states, transitions, and responses — are the brain's mechanism for efficient inference. Priors that work get reinforced; priors that fail get updated; over developmental and adaptive timescales, the generative model converges on a finite repertoire of cached priors that handles the range of situations the organism actually encounters.

Compensatory-defensive speech, we argued in Paper I, is a form of active inference under a specific architectural condition. It is the organism's pre-emptive response to the *predicted* error between the rigid social-normative prior (the interiorized self-as-should-be) and the current situation. The defense is a policy, selected to minimize that predicted error. Because the social-normative prior is protected from updating — by its developmental origin in caregiver authority, by ongoing social reinforcement, and by its high position in the hierarchy — the same class of policies gets repeatedly selected, reinforced, and cached. Under this account, the finite operational basis we will argue for is not an external imposition on the data; it is *what the predictive-processing framework specifically predicts for an agent organized around a protected prior with cached pre-emptive responses*.

This is the critical move for the present paper. The argument for a finite basis is not, or not only, an empirical generalization from clinical observation. It is a theoretical consequence of the framework within which the series as a whole operates. If defense is pre-emption of predicted shame, if that pre-emption is implemented as policy selection, and if policies with successful free-energy minimization get cached — then compensatory-defensive speech rests on a finite cached-policy inventory as a matter of mechanism, not as a matter of accident. The four arguments developed in §3 can be read, at a second pass, as four angles on the same predictive-processing prediction: the metabolic argument specifies why caching is enforced, the clinical argument shows what the cached inventory looks like from outside, the predictive-computational argument names the mechanism, and the developmental argument traces how the inventory is assembled across the lifespan.

2.3 Three candidate framings

We can situate the question of computability within three framings that have been offered, explicitly or implicitly, for natural language. Each has a distinguished history; each captures something real; each, we argue, is the wrong frame for our narrow domain. The three framings concern the *theory of language as computable object*; the predictive-processing framework of §2.2 specifies the *agent within which language is produced*. The two questions are separable, and we take positions on both.

The first framing is *radical inexhaustibility*. On this view, associated broadly with Romantic and post-structuralist traditions but still live in much of contemporary literary theory and continental philosophy of language, speech is fundamentally creative. Every utterance is a novel event; meaning arises in the ungovernable encounter between speaker, hearer, history, and world; any attempt to specify a finite basis misses precisely what makes language language. On this framing, formalization is not merely difficult — it is category-mistaken. We agree that this framing captures something important about certain domains of language: poetic speech, philosophical argument, living metaphor, the semantic creativity of a child acquiring her first language. We do not agree that it captures everything, and we deny that it captures our domain.

The second is *generative grammar*, associated with Chomsky (1957, 1965) and the computational-linguistic traditions that grew from it. On this view, syntactic competence rests on a finite set of rules capable of generating an infinite set of grammatical sentences. The Chomskyan program demonstrated that finite-basis formalization is possible for at least one substantial component of language — syntax — and this demonstration remains, in our view, one of the great achievements of twentieth-century cognitive science. But generative grammar deliberately brackets semantics, pragmatics, and — crucially for our purposes — the biological and affective substrate of speech. It tells us that *form* is tractable; it is explicitly silent about whether *function* is tractable in the same way.

The third is what we will call *constrained combinatoriality*: the view that a specific domain of a complex system can admit finite-basis formalization when that domain is bounded by a generative principle that is itself finite. This is the architecture of DNA, where four nucleotides composed by biochemical rules generate the entire space of biological forms; of chess, where a handful of piece-move rules generate a space that exceeds the number of atoms in the observable universe but remains, in principle, exhaustively specifiable; of phonology, where a finite segment inventory combines by a finite rule set to yield the sound systems of all human languages. Constrained combinatoriality does not claim that the generated space is small. It claims that the generating basis is.

The predictive-processing framework developed in §2.2 gives us the *mechanism* behind this general principle, in the specific case of compensatory-defensive speech. An agent organized around a protected prior implements its pre-emptive responses as cached policies, and that cached-policy inventory is the finite basis whose existence this paper defends. Constrained combinatoriality is not merely an analogy we are drawing to DNA and chess; it is what predictive processing specifically predicts for an architecture of this kind. The DNA and chess analogies are helpful intuition pumps — they illustrate how a small generative basis can produce an astronomical output

space — but they are not the argument. The argument is that the mechanism of policy caching in an agent with a protected prior entails a finite operational basis.

We propose, accordingly, that compensatory-defensive speech is a domain for which the third framing is correct, with the predictive-processing framework supplying the reason why. The four defensive vectors are finite. The pre-emption of predicted shame is a finite mechanism. The ecological/compensatory distinction is a finite criterion. A speech phenomenon organized by finite biological structures and implemented through cached-policy inference ought, by the mechanics of the framework, to inherit their finiteness at the level of description we will specify.

2.4 Speech as a privileged observable of active inference

Within the predictive-processing framework, most latent dynamics of active inference are not directly observable. We cannot measure prior precision, likelihood precision, or policy selection directly without invasive neural recording; we cannot watch the generative model update in real time. The framework describes a cognitive architecture whose fundamental variables are, for the most part, only indirectly accessible.

Speech is a specific and important exception.

An utterance is, within the predictive-processing framework, simultaneously three things. It is an *action*: a motor output selected from the organism's policy repertoire to sample the social environment, elicit responses, and test hypotheses about the interlocutor's state. It is a *sensory surface*: the speaker monitors their own output, and the coherence, fluency, and affective qualities of that output feed back as evidence for the speaker's own internal modeling of self-in-situation. And it is a *blanket state* in the sense of Friston's Markov-blanket formulation (Friston, 2019; Ramstead et al., 2018): the utterance constitutes the porous boundary between two inferring agents, the interface through which dyadic mutual inference occurs.

This triple status makes speech uniquely informative for the empirical study of active inference in social cognition. Neuroimaging captures neural correlates of inference, not the inference itself. Behavioral observation captures motor output, not its generative logic. Speech captures both: the active sampling strategy and the structure of the underlying generative model are both legible in the same observable stream. A speaker who insists "everything is fine" while exhibiting elevated threat-monitoring vocabulary demonstrates a specific precision-misallocation pattern — high confidence on a prior ("I am fine") that contradicts sensory evidence (affective distress signals). This is not a metaphorical application of the framework. It is the framework, observed through the medium that makes it most legible.

The implication for the present paper is consequential. If speech is a privileged observable of active inference, then compensatory-defensive speech — a specific and theoretically bounded subtype — is a privileged observable of a specific and theoretically bounded class of active-inference configurations. This is what makes its formalization both necessary and possible. Necessary, because the phenomenon is observable and important, and lacking formalization means the field describes it without mechanism. Possible, because the underlying dynamics are enumerable by the principles of the framework itself — not by fiat of the analyst, but by the structure

of the generative model that produces the speech. The question Paper II asks is not whether to impose a formalism on a recalcitrant object; it is whether to articulate the formalism that the underlying architecture has already implicitly defined.

2.5 Contribution and roadmap

The contribution of Paper II is to formulate the computability claim precisely within the predictive-processing framework, to marshal the convergent arguments that support it, to specify the alphabet of operations on which the formalization rests, and to characterize the four precision regimes that the operations trace as emergent attractors. We do not, in this paper, construct the annotation ontology or the classification engine; those are the work of Paper III. We do, however, provide what Paper III requires as input: a theoretically grounded list of operations, a characterization of the regime structure within which those operations cluster, a specification of the criteria by which ecological and compensatory expression are to be discriminated at the operation level, and the explicit methodological argument that this hypothesis space — being finite and theoretically principled — authorizes the ontology-bounded classification that Paper III operationalizes.

Section 3 develops four arguments for a finite operational basis: metabolic, clinical, predictive-computational, and developmental. Section 4 articulates the alphabet itself: four vector-organized families of speech-level operations, with examples and with explicit terminological boundaries against *templates* (which they are not) and *regimes* (into which they compose). Section 5 presents the four precision regimes — DEFENCE, COGNITIVE, ILLUSION, META — as emergent attractors within the predictive-processing architecture; the formal mathematical characterization of these regimes, together with a linear-Gaussian toy model and falsifiable predictions, is provided in Appendix A. Section 6 treats composition and context. Section 7 develops the ecological/compensatory discriminator at the operation level as a continuous, multi-component scalar. Section 8 states limitations and scope, makes explicit the three levels of description that the series invokes (biological vectors, precision regimes, engine-level attractors), disambiguates the shared terminology across those levels, and specifies the methodological bridge to Paper III. Section 9 concludes.

3. Four arguments for a finite operational basis

The claim that compensatory-defensive speech rests on a finite basis of operations can, at first hearing, seem either trivially true (of course there are only so many ways to be defensive) or implausibly reductive (but surely any utterance, including any defensive utterance, is infinitely varied in its detail). We argue that neither reading is right, and that the correct formulation — *the operations are finite; the sequences they generate are not* — is supported by four independent lines of evidence, each of which comes from a different intellectual tradition and converges on the same structural conclusion.

3.1 The metabolic argument

The first argument proceeds from a consideration of metabolic cost, and we should acknowledge at the outset that it is easy to over-read. The general principle — *sustained behaviors in predictive systems tend toward cached implementations* — is nearly a truism of any energy-constrained organism. Every reasonably frequent behavior, from walking to typing to ordinary conversation, runs on routines rather than on moment-to-moment assembly. If that were all we were claiming, the argument would prove something for all sustained behavior and therefore nothing distinctive for defense.

What makes the defensive case actually tight, and what gives this argument its real force, is not the general principle but the *intersection of three specific conditions*, all of which hold for sustained compensatory-defensive speech and none of which hold for ordinary speech.

The first condition is *elevated base cost*. Ordinary speech production rides on baseline neural activity. Defensive speech production rides on *mobilized* defensive physiology — sympathetic activation in fight and flight, ventral-vagal-with-affiliative-override in fawn, dorsal-vagal substrate in freeze. Mobilization, across all four vectors, is by evolutionary design a high-cost state. The autonomic, endocrine, and immune signatures reviewed in Paper I §6 do not come free; they are expensive because they were selected for brief, acute engagements in which energetic expenditure was traded for survival in an immediate emergency (Porges, 2011; McEwen, 1998).

The second condition is *absence of release*. Acute defensive responses were designed to fire, resolve, and then return to baseline. Compensatory defense, as we argued in Paper I §5, is pre-emptive response to *predicted* shame against a rigid social-normative prior — and the predicted threat does not resolve, because the rigid prior does not update. The social world which generates shame-prediction is present in nearly every interaction. There is no clear external signal that permits the defensive physiology to release. The organism is running a response designed for episodic engagement in a chronic mode.

The third condition is *continuous operation*. Language production, internal or external, is among the most temporally dense activities humans engage in. A person who converses for several hours per day and runs internal monologue for most of the remainder is producing speech almost continuously during waking life. If that speech is operating in the compensatory-defensive regime — if the defensive register is dominant — then the supporting defensive physiology is running not merely for the duration of episodic engagements but across the temporal span of the day.

The intersection of these three conditions — elevated base cost, absence of release, continuous operation — produces selection pressure on the implementation of compensatory-defensive speech that is substantially tighter than the pressure on ordinary speech. Ordinary speech is not produced on a mobilized substrate, gets plenty of release intervals, and runs on physiology that is relatively inexpensive at baseline. Compensatory-defensive speech has none of these reliefs. It runs on an expensive substrate, does not release, and runs continuously. The result is a metabolic regime in which inefficiency of implementation becomes, over developmental and life-span timescales, physiologically non-viable.

The conclusion we draw from this intersection is therefore not merely that compensatory-defensive speech *tends toward* cached operations — the general principle we started from — but something stronger: that it *must be* supported by a cached operational basis, and that the operations composing the basis must be *specifically tuned to the defensive task* rather than being generic cognitive routines recruited ad hoc. Generic routines would be less efficient than specialized ones, and under the selection pressure described, even small efficiency gains accumulate over developmental time into the difference between physiological viability and collapse.

We do not need, for the purpose of this argument, to specify the neural or psycholinguistic substrate of caching. Candidates include long-term potentiation of specific syntactic-pragmatic constructions, sedimented associative chains between shame-prediction signals and pre-fabricated response classes, and attractor dynamics in large-scale cortical networks. The argument does not rest on any of these. It rests on the logical structure of the metabolic constraint itself. Whatever the mechanism of caching turns out to be, *something* must serve that function, because the phenomenon as described in Paper I cannot otherwise exist in the form and duration in which it empirically does.

Within the predictive-processing framework adopted in §2.2, this argument specifies what kind of thing the caching mechanism must be. Caching, in the framework, is policy reinforcement under free-energy minimization; metabolic pressure is the selection force that enforces efficient policy retention. The metabolic argument is therefore not a separate consideration but the first face of the same predictive-processing prediction that the next three arguments will approach from different angles.

3.2 The clinical argument: stereotypy as evidence

The second argument comes from a direction that is easy to underweight if one has spent too long in theoretical neighborhoods and not enough in clinical ones. A hundred years of systematic clinical observation — across psychoanalytic, object-relations, attachment, social-cognitive, family-systems, and trauma-clinical traditions — has documented, across continents and schools and theoretical commitments, the remarkable *stereotypy* of defensive patterns.

That word — stereotypy — is, however, easy to invoke and hard to substantiate. For the clinical argument to do real work, we need to show that the convergence across traditions is genuine: that when a Kleinian analyst describes projective identification, a cognitive-behavioral researcher describes hostile attribution bias, and an attachment coder describes the linguistic signature of preoccupied anger, they are describing — at the level of observable speech and behavior — *the same phenomenon*, even if their theoretical vocabularies diverge sharply on what that phenomenon *means*. We offer four such cross-tradition triangulations, one per defensive vector family, as evidence.

Fight-family convergence: attributive hostility. Klein (1946) and Bion (1962) described *projective identification*, in which aspects of the patient's own unbearable experience are attributed to the interlocutor, who is then encountered as containing those qualities. Social-cognitive research independently developed, in a different vocabulary, the construct of *hostile attribution bias* (Dodge, 1980; Crick & Dodge, 1994),

documenting in experimental tasks the tendency of certain populations to read neutral or ambiguous behavior as containing hostile intent. Adult attachment research, using the Adult Attachment Interview's coding system (Main & Goldwyn, 1998; Hesse, 2016), identified a specific linguistic signature of "angry/preoccupied" speech about past caregivers, marked precisely by unsubstantiated attributions of negative intent. Three traditions, three vocabularies, one recognizable behavioral pattern: the attribution of hostile internality to the other in the absence of warrant.

Fawn-family convergence: compensatory self-erasure. Winnicott (1960) described the *false self*, in which the child's self-presentation is built to meet the caregiver's needs rather than to express the child's own. Bowen (1978) described *codependency* in family systems, with its characteristic pattern of self-effacement in service of relational maintenance. Ainsworth and her successors developed the attachment category initially termed "anxious-resistant" and later refined to "anxious-preoccupied" (Hazan & Shaver, 1987; Mikulincer & Shaver, 2016), in which adult relational behavior is organized around *hyperactivating strategies* — accommodation, proximity-seeking through alignment, suppression of own need-expression. Walker (2013) independently described *fawn* as a complex-trauma response, with its characteristic linguistic markers of ingratiation and self-minimization. Four traditions — object-relations, family-systems, attachment, trauma-clinical — converge on the same speech-level phenomenon: systematic erasure of the speaker's own position in anticipated service of the interlocutor.

Flight-family convergence: abstraction as distance. Anna Freud (1936) named *intellectualization* as the use of abstract reasoning to avoid emotional contact with threatening content. Bowlby (1980) described *defensive exclusion*, the systematic barring of affective material from representation. Attachment research identified the characteristic *deactivating strategy* of dismissing-avoidant adults, with its linguistic signature of affect-light, generalized, history-poor narrative when the Adult Attachment Interview probes attachment-relevant material (Main & Goldwyn, 1998). The *alexithymia* construct (Sifneos, 1973; Taylor et al., 1997) documented difficulty translating affective experience into specific linguistic content. Fonagy and colleagues (2002) described *pseudo-mentalization*, the production of reflective-sounding language without affective grounding. Five traditions, five vocabularies, one recognizable pattern: response to affectively-live material by lifting into abstract or generic register.

Freeze-family convergence: epistemic absence. Schauer and Elbert (2010) documented *dissociative speech* in trauma survivors, characterized by disrupted narrative coherence and loss of first-person agency. Main and Hesse (1990) identified *unresolved/disorganized* attachment through its AAI signature of characteristic lapses in monitoring of discourse — speech that breaks down, becomes vague, or loses referential anchoring specifically when trauma-relevant material is probed. Van der Kolk (2014) synthesized across these traditions the behavioral phenomenology of freeze responses in survivors of developmental trauma. Three traditions converge on the same observable pattern: speech that declines to constitute a coherent, evaluable first-person position when such a position would be called for.

We want to mark, at this point, a distinction the argument requires. The claim we are making is about convergence at the *observational* level — what the different traditions record as happening in the speech and behavior of their subjects. We are

not claiming that the traditions converge on what this behavior *means* theoretically, and in most cases they visibly do not. A Kleinian holds that projective identification involves an unconscious phantasy of expulsion and containment; a cognitive-behavioral researcher holds that hostile attribution bias is a learned information-processing pattern; an attachment coder holds that the linguistic signature reflects internal working models of relationships. These are *different* theoretical commitments, and we take no position on which of them is right. For the computability argument, what matters is that trained observers from these different theoretical traditions, looking at the same speech samples, would agree on *what is happening* even when they disagree on *why*. The observational convergence is what gives us traction; the theoretical divergence is what keeps the field honest about its unsolved problems.

The conclusion is therefore the same one we stated at the opening, but now with content: the stereotypy of compensatory-defensive speech across clinical traditions is not a vague impression. It is a documented convergence of independent observational systems on a small number of recognizable patterns. If the phenomenon were genuinely unbounded, no such convergence would be possible. That the convergence exists, in the form it does, across the traditions it does, is the empirical basis on which the finite-operational-basis claim stands.

Read through the predictive-processing framework of §2.2, this convergence has a clean explanation: independent observational systems converge on the same recurring patterns because those patterns are the external footprint of a shared architectural process — cached-policy deployment against predicted shame-error, implemented across individuals with different histories but facing structurally similar predictive demands. The clinical literature is, on this reading, the most extensive naturalistic record of predictive-processing dynamics we currently have, collected across a century by observers who were not, for the most part, thinking about predictive processing at all.

3.3 The predictive-computational argument

The third argument names the mechanism that the first two imply. If compensatory-defensive speech runs on a cached operational basis (§3.1) and if that basis is detectable across independent clinical traditions (§3.2), the obvious question is: *what process builds and maintains this cache*? The predictive-processing framework adopted in §2.2 answers this question with unusual specificity, and this specificity is what makes the third argument more than a restatement of the first two.

Within the framework, the brain continuously predicts its own sensory and affective states (Clark, 2013; Hohwy, 2013; Seth, 2021). Shame, as Paper I argued, arises as prediction error between the rigid social-normative prior and the currently perceived self-in-situation. But this prediction-error signal is itself generated by a prediction: the system predicts, ahead of the fact, whether the current trajectory of self-presentation will produce a shame-error or will not. When that *predicted* shame-error crosses a threshold, the system generates a pre-emptive response — a defensive speech act — to alter the trajectory before the error consolidates.

This places defensive speech at a specific architectural location: it is *inference about inference*, or more precisely, it is policy selection over the agent's own expected

prediction-error trajectory. The defensive speech act is not, at the deepest level, a response to reality; it is a response to the system's own prediction about how the trajectory is about to unfold relative to a protected prior. The consequence is that defensive policies can be evaluated *only* against their predicted effect on the agent's own internal signal, not primarily against external outcomes. A defensive move "works," from the system's own perspective, when it prevents or attenuates the predicted shame signal, regardless of its external effects on the interlocutor or on the situation at hand. This asymmetry — that the defensive move's success criterion is internal, not external — is, we will argue in §5, what produces the characteristic precision biases that compensatory-defensive speech exhibits.

This evaluative structure is what produces caching. A policy that reliably attenuates predicted shame in a particular class of situations will be reinforced every time it succeeds. A policy that fails will be deprecated. Over developmental time, the system converges on a small inventory of policies, each tuned to a specific class of shame-predictions, each selected because of its reliable attenuating effect. This convergence is not a side effect of the framework — it is what the framework predicts for any system implementing pre-emptive defense against a protected prior.

The inventory that results has structural properties we can specify. First, it is *vector-organized*: the four biological defensive vectors from Paper I provide four distinct mechanisms for attenuating predicted shame (attack the source, withdraw from it, appease it, dissociate from it), and policies within each vector family share mechanism-level similarities that distinguish them from policies in other vectors. Second, it is *finite*: the number of mechanism-distinct policies is bounded by the structure of the defensive architecture itself, not by enumerative convenience. Third, it is *composable*: single situations often activate multiple vectors, and policies from different families can be deployed in sequence or in combination within a single utterance — a composability we treat more fully in §6. Fourth, and importantly for §5, it is *characterized by precision biases*: each policy operates by adjusting the balance between prior precision and likelihood precision in a specific way, and these adjustments aggregate, across the lifetime of an agent, into the four precision regimes we develop in §5.

The connection to the rest of the series becomes direct at this point. If defensive policies are cached in a bounded, mechanism-organized inventory, then the policies have structured observable signatures — specific linguistic patterns that are the policies' outputs. Those signatures are what the engine of Paper III is designed to detect: a classification pipeline whose hypothesis space is the cached defensive inventory, operating on the speech that expresses it. The methodological warrant for ontology-bounded classification, which this paper sets out to provide, rests on precisely the prediction the predictive-processing framework makes about an architecture of this kind.

We do not, in this argument, claim more than the framework licenses. We do not claim to have modeled the specific neural implementation of policy caching. We do not claim that the framework is the only description under which the finite-basis claim is defensible; the metabolic and clinical arguments stand on their own. What we claim is narrower: *given* the predictive-processing framework, compensatory-defensive speech as characterized in Paper I is *predicted* to rest on a finite cached-policy inventory

with the four structural properties above. That prediction converges with what §3.1 and §3.2 independently find. The convergence is what gives the overall argument its force.

3.4 The developmental argument

The fourth argument comes from developmental and attachment research, which has established — across longitudinal studies, clinical observation, and cross-cultural comparison — that defensive patterns *crystallize*. The child who develops an avoidant attachment pattern does not improvise a novel avoidance strategy on each new occasion; by mid-childhood, she has a recognizable repertoire of moves that she deploys repeatedly, with minor variations adapted to context, across the full range of her relational life. The same is true for anxious-preoccupied patterns, disorganized patterns, and the fawn-configurations that Walker (2013) documented in complex developmental trauma. These patterns are not loose dispositions; they are specific and reproducible.

The developmental research most directly relevant to our argument comes from Mikulincer and Shaver's (2016) comprehensive review of adult attachment patterns in speech and relational behavior. Their meta-analytic work demonstrates that attachment-pattern-linked linguistic and behavioral signatures are stable across decades, consistent across raters, and predictive of downstream outcomes. Fonagy and colleagues' (2002) work on the reflective-function scale similarly shows that failures of mentalization take recognizable, recurring forms — and those forms cluster into a small number of types rather than distributing continuously across a boundless space.

The developmental argument is therefore this: defenses are *learned*, and what is learned tends to crystallize into a repertoire. This is true across domains — motor skills, social scripts, expert judgment — and there is no reason to think defensive speech is exceptional. The compensatory-defensive speech that a given adult produces today is not freshly invented; it is the sedimented, consolidated outcome of thousands of developmental moments in which specific defensive moves either succeeded (and were reinforced) or failed (and were deprecated). The adult speaker is not creating the repertoire. She is deploying it.

Within the predictive-processing framework, this developmental trajectory is the assembly process through which the cached-policy inventory of §3.3 is built. Each developmental moment in which a defensive move successfully attenuated a predicted shame-signal, or failed to do so, is a data point in the lifetime-scale policy reinforcement that produces the adult repertoire. The stability and cross-context consistency that attachment research documents at the behavioral level are the observable trace of this assembly: by adulthood, the inventory is stable because it has been extensively selected, and stable inventories produce consistent outputs across situations. The developmental argument thus specifies the temporal dimension of the predictive-processing argument — how the cached inventory comes to exist in the form in which §3.2's clinical observers encounter it.

3.5 Convergence and what it implies

Four arguments, drawn from four traditions — physiology, clinical observation, computational neuroscience, developmental psychology — converge on the same structural conclusion: compensatory-defensive speech rests on a finite, cached, re-used, crystallized operational basis. The convergence is not accidental. Each argument can be read, as §2.2 anticipated, as a different angle on a single underlying prediction of the predictive-processing framework: that an agent organized around a protected prior, implementing pre-emptive defense against predicted prediction-error in that prior, should accumulate a finite, mechanism-organized policy inventory over developmental time. Physiology (§3.1) provides the selection pressure that enforces efficiency. Clinical observation (§3.2) provides the external record of the inventory's contents. The predictive-computational argument (§3.3) specifies the mechanism by which the inventory is built and maintained. Developmental research (§3.4) provides the assembly process through which the inventory is constructed. Each argument alone might be contested; their convergence, and their mutual fit with the predictive-processing framework, makes the contest substantially harder. It is not merely possible that the operational basis is finite; it is, on the weight of evidence, unlikely to be otherwise.

The immediate implication is that the computability question has an affirmative answer at the level of the operational basis itself. If the basis is finite, it can be enumerated; if it can be enumerated, it can be labeled; if it can be labeled, it can be recognized by trained human annotators or by computational classifiers. Formalization is not, on this view, a hope or an aspiration. It is a direct consequence of what compensatory-defensive speech structurally is.

What formalization cannot do, and what we are careful not to claim, is to exhaust the *products* of the operations. The sequences of compensatory-defensive speech that any given speaker produces in any given life are as various as lives are various. But variety at the sequence level is entirely compatible with finiteness at the operational level, as the DNA and chess analogies make clear. Four nucleotides generate the biosphere; forty-ish opening choices, a handful of piece-move rules, and a sixty-four-square board generate ten-to-the-hundred-and-twenty chess games. The finiteness lives in the generating basis, not in what it generates.

4. The alphabet: operations, not templates

Having argued that compensatory-defensive speech rests on a finite operational basis, we now specify what that basis contains. This section introduces the alphabet itself: a vector-organized inventory of speech-level operations. Before listing them, we fix terminology, because the field around defensive speech is littered with near-synonyms and the precision of the formalization depends on our distinguishing them cleanly.

4.1 Operations, templates, regimes

A **template** is a fixed linguistic form — a stock phrase, a cliché, a pre-fabricated sentence that is deployed as-is. Templates exist in language (greetings, condolences, some formulae of politeness), and a small subset of defensive speech does instantiate

them. But templates are not what we mean by operations, and the conflation of the two would be, we think, the fastest way to discredit the entire combinatorial claim. If operations were templates, they would be too rigid to explain the observed variety of defensive speech; critics would (rightly) point out that real defensive speech is generative, not cut-and-paste. We agree, and we propose operations precisely because they are not templates.

An **operation**, in our sense, is a *speech-level generative move*: a rule for transforming an input context into an output utterance that performs a specific defensive function. A useful analogy is the chess move. A knight does not have a "knight-template"; it has a *rule of movement* — two squares in one direction and one perpendicular. From that rule, combined with the current state of the board, any specific knight move is generated. Two knight moves from two different positions share the same underlying rule but produce entirely different concrete moves. So with operations: "you-attribution" is a rule for transforming an internal experience the speaker is unable to acknowledge into a second-person declaration about the interlocutor, and the concrete utterances generated by applying that rule in different contexts vary indefinitely while preserving the rule.

A **regime** is a *sustained configuration of operations* that a speaker adopts across a stretch of discourse. Where an operation is the individual move, a regime is the strategy within which moves are selected. A speaker who is chronically in a projective regime deploys predominantly fight-family operations, with characteristic co-selection patterns and characteristic avoidance of operations from other families. Regimes are thus what the clinical tradition has generally called "defensive patterns" or "defensive styles" — and they are, in our architecture, *emergent from operation-selection probabilities* rather than primitive themselves. Paper IV will develop the formalization of regimes and their trajectories; here we name them only to distinguish them from operations, which are the lower-level, more directly computable primitive.

The relation among the three levels is thus: **templates < operations < regimes**. Templates are not our primitive because they are too rigid; regimes are not our primitive because they are emergent; operations are the right level of description because they are generative (unlike templates) and directly observable in single utterances (unlike regimes, which require longer stretches of discourse to become visible).

4.2 Fight operations

The fight vector preserves its outward-directed logic when it becomes speech, and the family of operations through which it does so shares the signature of directing attributive or evaluative force toward the interlocutor or a represented third party. We list here the core operations of the fight family, with the understanding that this list is provisional — Paper III will refine it through operationalization and testing.

You-attribution. The speaker makes a declarative claim about the interlocutor's internal state, motivation, or character, in a form that presupposes privileged access to that state. Examples: "you don't care"; "you're trying to control me"; "you think you're better than me". The operation is to convert an unresolved internal prediction ("my needs are being disregarded") into a second-person declarative about the other.

The key marker is the assertion of interlocutor internality in the absence of epistemic warrant.

Global-quantifier deployment. The speaker characterizes the interlocutor's behavior with always/never quantifiers that exceed the scope of the current exchange. Examples: "you always do this"; "you never listen"; "this is what you always turn every conversation into". The operation converts a present-tense grievance into a claim about pattern, and in so doing removes the possibility of contextual repair.

Externalized-agency framing. The speaker recounts an interaction in a grammatical construction that absents their own contribution from the account. Examples: "so then you started attacking me"; "and then you made me upset". The operation systematically places the interlocutor in causal-agent position while placing the speaker in recipient-of-effect position, irrespective of what the interaction structurally was.

Motive-attribution. The speaker declares not just what the interlocutor did but why, in terms that the interlocutor has not granted. Examples: "you said that just to hurt me"; "the real reason you brought this up is…" The operation is close to you-attribution but specifically concerns the causal-intentional reading of a concrete act, rather than a general claim about interlocutor character.

Each of these operations has its ecological counterpart. "I notice you haven't responded to my last three messages; it makes me wonder whether I've misread something" is not a fight operation; it is a proportionate first-person inquiry that preserves epistemic humility. The operation-level family is the same underlying vector; the ecological/compensatory distinction, which we develop in §6, cuts across the family at a level that specifies how the operation is calibrated to context.

4.3 Flight operations

The flight vector preserves its retreat-directed logic when it becomes speech. The family of flight operations shares the signature of producing semantic distance between the speaker and a topic that would, if engaged, constitute contact.

Abstract-register shift. The speaker responds to a concrete, affectively-live prompt with a response pitched at a general or theoretical register. Example: asked "how are you feeling about your mother's illness?", the speaker responds "it's interesting how we construct meaning around mortality". The operation converts a first-person affective inquiry into a third-person abstract reflection, preserving the topic nominally while removing the self from the topic.

Topic-drift. The speaker engages a prompt partially and then shifts the referential target before the original prompt is resolved. Example: asked about a specific incident, the speaker begins a response that references the incident but then migrates to a different, related-but-safer topic. The operation is recognizable by the characteristic failure to return to the original prompt even when invited.

Meta-commentary replacement. The speaker substitutes commentary about the conversation for continued participation in it. Example: "this is getting intense"; "I notice we always end up here"; "isn't it interesting that we keep returning to this". The operation removes the speaker from the first-order interaction by elevating her into the

position of observer-of-the-interaction.

Tense or modal displacement. The speaker responds to a present-tense affective prompt in the past tense, the conditional, or a hypothetical. Example: asked "what do you want right now?", the speaker responds "well, I think what someone in my position would typically want is…" The operation distances the self in time or modality from the affective live wire of the question.

As with the fight family, each of these operations has an ecological counterpart. "Let's pause this conversation; I need to think, and I want to come back to it tomorrow" is a fight for flight's vector deployed ecologically. It is named as retreat, proportionate to the situation, and preserves the conditions for return.

4.4 Fawn operations

The fawn vector preserves its submit-to-appease logic when it becomes speech. The family of fawn operations shares the signature of minimizing the speaker's own contour while amplifying accommodation of the interlocutor's assumed preferences.

Hedge-stacking. The speaker places multiple softeners around a claim that does not require them, such that the claim's assertoric force is drained before it is made. Example: "I was just kind of maybe wondering if perhaps it might sort of be…" The operation produces a statement whose syntactic shape is preserved but whose pragmatic force is effectively zero, reducing the speaker's exposure.

Apology surplus. The speaker issues apologies or apology-markers in contexts where no transgression has occurred and none has been signaled. Example: "sorry, I know this is a stupid question"; "sorry to bother you with this"; "I apologize for taking up your time". The operation converts a neutral interactive moment into a minor submission ritual.

First-person-agency erasure. The speaker expresses preferences, disagreements, or desires in grammatical constructions that attribute them to generic others or to no one. Example: "people might say…"; "someone in this situation could feel…"; "one does start to wonder whether…" The operation removes the speaker's first-person presence from claims that would otherwise carry it.

Anticipatory agreement. The speaker aligns with a position the interlocutor has not yet taken, pre-empting possible disagreement with agreement. Example: "you're probably going to say X, and I totally see where you're coming from on that". The operation collapses the interactive space in which genuine disagreement could emerge before it emerges.

Upward-grooming insertion. The speaker inserts compliments, affirmations, or expressions of appreciation at points that lack compensatory content. Example: responding to a neutral question with "you always ask such thoughtful questions" before providing the answer. The operation performs linguistic grooming without being triggered by the interaction itself.

The ecological counterparts here are particularly important to keep in view, because fawn is the defensive family most easily moralized. The junior clinician who says to a senior supervisor, "I defer to your judgment on the differential — you've seen

hundreds of these, and I've seen one", is deploying the fawn vector ecologically: the hierarchy is real, the deference is proportionate, and the speaker's own perspective is not erased.

4.5 Freeze operations

The freeze vector preserves its immobilization-as-disappearance logic when it becomes speech. The family of freeze operations shares the signature of producing speech that resists being evaluable — that declines to constitute a position the interlocutor can engage with.

Epistemic vagueness. The speaker deploys epistemic hedges ("I guess", "I suppose", "kind of", "I don't really know") in contexts where access to the relevant content would ordinarily be available. Example: asked what she wants for dinner, the speaker responds "I don't know, I guess whatever". The operation specifically targets epistemic assertion — the speaker's claim to know her own state or preference — and attenuates it.

Deictic absence. The speaker uses vague or generic reference where specific reference would be more informative. Example: "things are kind of hard right now"; "it's been a lot". The operation evacuates deictic anchors (specific persons, times, events, emotional qualities) from a statement that nominally concerns them.

Reference-breakage. The speaker uses pronouns or descriptions that fail to pick out determinate referents. Example: "they said something, and then this happened, and now it's weird". The operation produces speech whose surface form is grammatical but whose propositional content is indeterminate.

First-person-commitment absence. The speaker makes claims about her own states in constructions that remove her from the claim. Example: "things feel off"; "there's something weird going on"; "it's been hard". The operation removes the I from a first-person claim, producing something that is structurally a description of a state without being, quite, the speaker's own.

Chronic indefinite-response. The speaker, when invited to specify, produces repeated indefinite or minimal responses. Example: asked to elaborate on "I don't know, it's been hard", the speaker responds "yeah, just, like, I don't know, it's just like that". The operation is recognizable by its survival of repeated invitations to specify.

The ecological counterparts of freeze operations are genuine not-knowing and receptive silence. "I honestly don't know yet; let me think about it and come back to you" is the freeze vector deployed ecologically. The speaker acknowledges her present limitation, remains in contact with the conversation, and holds open the possibility of return to substance.

4.6 The alphabet as a whole

The four families together — fight, flight, fawn, freeze — yield an operational alphabet of on the order of twenty core operations, with internal sub-variations that extend the count somewhat. This is the basis on which Paper III will construct the annotation ontology. Several features of this alphabet are worth noting before we proceed.

First, the operations are *not mutually exclusive* within an utterance. A single compensatory-defensive utterance can instantiate more than one operation simultaneously — for instance, a you-attribution delivered with hedge-stacking, or an abstract-register shift that is also a meta-commentary. The formalization will need to handle this, and Paper III does so by allowing multi-label assignment at the operation level.

Second, the families are *not mutually exclusive across a speaker's repertoire*. A speaker can deploy fight operations in one context and fawn operations in another, and many speakers do. The regime-level analysis of Paper IV captures the tendencies of operation-selection across time and context; the operation level itself is silent about regime.

Third, and most importantly, the alphabet is *not a diagnostic instrument*. An individual utterance that instantiates a compensatory-defensive operation is not, in itself, a marker of pathology, dysregulation, or any other evaluative category. Humans deploy these operations constantly, in every conversation, as part of ordinary speech production under ordinary social pressure. The pathological shape, where it exists, lives at the regime and trajectory levels (where does the speaker live across time? can she move?), not at the operation level. We emphasize this because the diagnostic mis-reading is a predictable failure mode of the formalization, and it is one we intend the engine of Paper III to structurally prevent.

5. Four precision regimes as emergent attractors

Up to this point the paper has developed two levels of description: the biological vectors of Paper I (fight, flight, fawn, freeze, preserved as directional logic in speech), and the speech-level operations of §4 (the bounded inventory of generative moves that instantiate those vectors in actual utterances). Both levels are necessary for what we are doing; neither is sufficient on its own. The vectors are too coarse to ground annotation: they tell us what directional logic an organism inherits from its defensive repertoire, but not how any specific utterance realizes that logic in linguistic form. The operations are too local to capture the sustained configurations that clinical observers recognize, and that any engine will need to detect across stretches of discourse rather than within single utterances. Between the vector and the operation lies a third level, which this section articulates: the *precision regime*.

A precision regime is a configuration of precision allocation within the hierarchical generative model of §2.2: how much weight is placed on prior expectations relative to incoming evidence, which channels of disconfirmatory information are attenuated or amplified, and with what confidence the system commits to its selected policies. The four regimes we articulate below — **DEFENCE**, **COGNITIVE**, **ILLUSION**, **META** — are not categories we impose on the data; they are attractor-like configurations in the precision-parameter space, states into which the system settles when certain combinations of priors, evidence, and policy incentives persist. The formal specification of these attractors — precision hyperparameters, a linear-Gaussian toy model, and five falsifiable predictions — is provided in Appendix A. In this section we characterize them conceptually, and show how each realizes, at the level of precision dynamics, the phenomena described at the operation and vector levels.

5.1 Why four?

The number four is not axiomatic; it is derived from two orthogonal axes of variation in precision control under hierarchical inference. The first axis is the *direction of precision bias* — whether the system places greater effective weight on protected priors (prior-dominated updating) or on incoming evidence (evidence-sensitive updating). The second axis is the *depth of inference* — whether precision parameters are treated as fixed at the current level of inference (first-order), or whether precision itself becomes an object of inference, subject to modulation and learning (higher-order).

The intersection of these two axes yields four natural quadrants. DEFENCE occupies the prior-dominated, first-order quadrant: protected priors dominate evidence, and this domination is not itself under reflective scrutiny. COGNITIVE occupies the balanced, first-order quadrant: evidence and prior are weighted context-sensitively, without higher-order modulation beyond what the task requires. ILLUSION occupies a related but distinct region of the prior-dominated space: the domination of evidence does not require acute threat to sustain itself, and the configuration is self-reinforcing rather than episodically threat-driven. META occupies the higher-order quadrant, orthogonal to the first three: precision itself becomes an inferential object, enabling the system to modulate the balance between prior and evidence as volatility and context change.

These are minimal quadrants in the sense that Bayesian model reduction could, in principle, find finer structure within them — ILLUSION might decompose into subtypes, META might have its own internal differentiation — but the four-regime architecture is not inherited as revelation. It is the simplest partition that preserves the distinctions the clinical literature documents while remaining theoretically principled within the framework. If a later structure-learning analysis finds that ILLUSION decomposes into two distinct attractors with different precision signatures, the framework accommodates that refinement without disturbing the underlying logic; what it does not accommodate is a reduction of the four to fewer, because the clinical convergence documented in §3.2 actively argues against such reduction.

5.2 DEFENCE

DEFENCE is the regime in which the four biological vectors of Paper I, when they become chronic and pre-emptive, find their precision-theoretic expression. It is characterized by elevated prior precision on self-protective priors — the interiorized social-normative prior against which shame is predicted — and by suppressed likelihood precision on the channels that would carry disconfirmatory evidence. The effect is that evidence contradicting the protected prior is systematically down-weighted in belief updating; the rigid prior remains intact because the mechanism by which it could be updated has been attenuated. Policy precision on defensive policies is also elevated: the system commits to them with unusual confidence, because the success criterion for a defensive policy (§3.3) is internal attenuation of predicted shame rather than external outcome, and internal attenuation is reliably achieved by the policies that have been cached for exactly this purpose.

The operations of §4 are the speech-level signatures of this precision configuration. A you-attribution (§4.2) reflects a precision regime in which the speaker's protective prior about her own state is weighted so heavily that evidence of the interlocutor's actual state cannot compete; the prior is projected outward, attached to the other, and defended there. An abstract-register shift (§4.3) is a policy that, by design, samples the conversational environment in a way that minimizes the flow of threat-relevant disconfirmatory evidence into the generative model. A hedge-stack (§4.4) is a pragmatic structure that drains the speaker's own assertoric exposure and, with it, the likelihood of generating the kind of evidence that would trigger shame-prediction. An epistemic-vagueness move (§4.5) attenuates first-person claims whose truth values would otherwise constitute evidence against the protected self-model.

A critical clarification: DEFENCE in this technical sense is *not* the same as acute defensive arousal in the face of a real, immediate threat. The latter is, in predictive-processing terms, entirely appropriate — a high prior precision on "this situation is dangerous" accompanied by reduced likelihood precision on reassuring but potentially deceptive cues is good Bayesian policy when the environment actually is threatening. DEFENCE, as we use the term, refers to the chronic, pre-emptive configuration in which the protected prior is not the immediate environment but the interiorized social-normative self-model, and the evidence being filtered is not reassurance from a potential enemy but everyday social input from an ordinary interlocutor. The architectural machinery is the same; what differs is the prior being protected, the context in which it is protected, and the duration over which protection is sustained.

5.3 COGNITIVE

COGNITIVE is the regime in which precision is allocated in a balanced, context-sensitive manner: prior precision and likelihood precision are set to levels that track the actual reliability of priors and evidence in the current situation, and policy precision is preserved at levels that permit cooperative repair, information-seeking, and responsive updating. It is the regime in which the ecological counterparts of the operations listed in §4 become legible: the first-person inquiry instead of you-attribution, the named and reversible retreat instead of abstract-register drift, the proportionate hedging instead of hedge-stacking, the acknowledged not-knowing instead of epistemic vagueness.

COGNITIVE should not be misread as "absence of defense." An agent in the COGNITIVE regime is still defending priors — every predictive system does; defense, in the broad sense, is the organism's basic mode of homeostatic self-maintenance. What distinguishes COGNITIVE from DEFENCE is not the presence or absence of prior-protection but the *calibration* of prior precision to the actual reliability of the prior against the actual reliability of available evidence. In COGNITIVE, a prior is weighted high when it is well-supported and its weight decreases when evidence accumulates against it. In DEFENCE, a prior is weighted high because it is protected — because the mechanism that would adjust its weight in response to evidence has been attenuated.

This distinction is what the ecological/compensatory discriminator of Paper I §3.5 tracks at the precision level. An ecological defensive move is deployed from within the

COGNITIVE regime: it engages the appropriate biological vector at a magnitude proportional to the present trigger, for the duration that the situation actually requires, and releases when the situation is resolved. A compensatory defensive move is deployed from within the DEFENCE regime: it engages regardless of current warrant, scaled to the protected prior rather than to the current threat, and persists because the protected prior does not update. The same biological vector — the same ordinary human capacity to push back, withdraw, appease, or go still — generates ecological expression in COGNITIVE and compensatory expression in DEFENCE. What differs is the regime in which the vector is deployed.

5.4 ILLUSION

ILLUSION occupies a distinct region of the prior-dominated configuration space, one which — importantly — can be sustained without the acute threat signals that drive DEFENCE. It is characterized by biased prior means (the prior is not merely weighted heavily; its central tendency is offset from the evidence-supported value) combined with selective reduction of likelihood precision specifically on those channels that would disambiguate the bias. Policy selection in ILLUSION systematically avoids epistemic actions that would generate disambiguating evidence: the speaker does not ask the question that would resolve the ambiguity, does not invite the feedback that would test the narrative, does not seek the perspective that would challenge the frame. The epistemic weight κ (Appendix A) is, in the relevant subspaces, structurally reduced.

The difference between DEFENCE and ILLUSION is the difference between *rigidity under pressure* and *coherent bias without pressure*. DEFENCE needs threat-prediction to sustain its configuration; when the threat is resolved, DEFENCE releases, returning the precision allocation toward its baseline. ILLUSION does not need threat-prediction; the configuration is self-reinforcing, drawing its stability from the internal coherence of the narrative itself rather than from any ongoing external pressure. This is why certain patterns of speech — sustained over years, deployed in situations that contain no acute threat, resistant to feedback not because the speaker feels attacked but because the frame renders the feedback invisible — require ILLUSION rather than DEFENCE to be captured.

Clinically, ILLUSION is what the psychoanalytic and attachment traditions have described under various labels: false self organizations that have stabilized across the lifespan, grandiosity that organizes perception rather than being provoked by slight, sustained narratives of relational history in which the speaker is consistently the protagonist of a story whose logic is never disturbed by the actual behavior of others. The computational point is that these phenomena share a specific precision signature: it is not that evidence is being actively rejected under pressure; it is that the system has settled into a configuration in which the evidence that would challenge the frame does not enter the generative model's updating machinery at all. The signature of ILLUSION at the speech level is accordingly not primarily the operations of acute pre-emption (hedge-stacking, abstract-register shift) but the longer-timescale patterns of narrative maintenance, selective topic selection, and consistent framing that Paper IV will examine in detail.

5.5 META

META is the regime in which precision itself becomes an inferential object. The system does not merely update beliefs about states of the world given precision-weighted evidence; it updates beliefs about the reliability of its own evidence-weighting. Formally (Appendix A), this is modeled by treating precision parameters as random variables subject to hyperpriors and higher-order prediction errors: when the system's predictions systematically fail, or systematically over-succeed, the prediction-error statistics themselves become evidence about whether the current precision allocation is well-calibrated. The epistemic weight κ is, in META, elevated precisely on those actions that would provide information about precision — the questions that test one's own framing, the invitations to feedback that probe one's own model of the interlocutor.

The phenomenology of META is what people recognize as reflective capacity, epistemic humility, or metacognitive awareness — the ability to notice that one's confidence in a prior is disproportionate to the evidence for it, that one's weighting of evidence may itself be miscalibrated, that a pattern of conversational failures suggests one's model of the interlocutor or of oneself needs revision. In speech, META is marked by operations that the other regimes systematically lack: the first-person epistemic qualifier that actually tracks uncertainty ("I don't know whether I'm reading this situation correctly" as opposed to the DEFENCE-regime hedge-stack that drains assertoric force without engaging uncertainty), the invitation to disconfirmation ("tell me if I'm getting this wrong"), the explicit reframing of one's own prior stance in light of what the interlocutor has just said.

META is the regime that makes *transitions between regimes* possible as an explicit, policy-governed process rather than as an unintended drift. A speaker who can reflect on her own defensive-speech patterns can, from within META, adjust the precision allocation that sustains DEFENCE or ILLUSION; a speaker who cannot access META is largely confined to the regime she currently inhabits, with transitions occurring — if at all — through external pressure rather than through internal modulation. This is the architectural correlate of what clinical traditions have called capacity for reflection or mentalization; in our framework, it is the regime whose activation is what makes therapeutic change mechanically possible. Paper IV will develop the trajectory-level consequences of this property.

A caution: META is easily misread homuncularly. It is tempting to imagine a "higher self" doing the reflecting. The formal account (Appendix A) makes explicit that META requires no such agent — only that precision parameters be treated as random variables within the same hierarchical inference that everything else is doing. META is continuous with the lower regimes, not ontologically separate from them. This non-homuncular framing is important not only philosophically but empirically: it means that META, like the other regimes, has observable speech-level signatures and can be detected by the same classificatory apparatus that detects the others.

5.6 Operations as the speech-level signatures of regimes

The relationship between the alphabet of §4 and the regimes of this section is the relationship between observable output and underlying dynamical pattern. Each operation has a characteristic precision signature: it is generated when certain precision parameters are configured in certain ways, and it produces, as a consequence of that configuration, a speech-level move with specific linguistic properties. A you-attribution is not just a semantic category; it is the speech-level signature of a precision configuration in which prior precision on self-protective priors is high and likelihood precision on interlocutor state is low. An abstract-register shift is not just a topic drift; it is the speech-level signature of a policy selected under conditions of suppressed epistemic value on affectively-live material. And so on for each operation in §4.

This is what licenses the inference from observable speech to underlying regime. An annotator — human or computational — who sees an utterance instantiating a specific operation can infer, with appropriate confidence, that the speaker is currently in a precision configuration consistent with that operation. An annotator who sees sustained co-selection of operations from the same family, across several utterances, can infer with higher confidence that the speaker is in the regime of which those operations are the characteristic outputs. The regime itself is not directly observable, but it is *inferable* from the pattern of operations, which are observable. This is the same epistemic structure that governs any attractor-based inference in complex-systems analysis: one observes the trajectories, one infers the attractor.

The methodological payoff for the engine of Paper III is specific. The engine does not need to directly detect precision configurations — a task that would require invasive neural recording and that no text-based classifier can perform. The engine needs only to detect operations, which are linguistically realized and bounded in number, and to track their co-occurrence patterns across time. The regimes emerge from the operations; the operations are what the engine works on. Paper II's claim that operations are finite and theoretically principled is, on this reading, exactly the claim that authorizes the engine's classification architecture: a finite observable basis (operations) gives empirical access to a finite latent structure (regimes) via the mapping that the precision framework specifies. A full account of how this mapping is realized in the Mindloom engine — including the important terminological point that Paper II's four precision regimes are realized in the engine's ontology as *layers of analysis*, while the engine's vocabulary of ten observable behavioral attractors (BUILD, SEEK, UNSEAL, LOCK, SEAL, DRAIN, FLOOD, EDGE, SHIFT, VOID) refers to a lower level of description — is given in §7.

6. Composition and context

Having characterized operations as the observable basis (§4) and regimes as the attractors they trace (§5), we now address the question that has been deferred in both: how do operations actually combine in real speech, and what role does context play in making those combinations legible? This section develops five structurally important points: how operations compose within a single utterance (§6.1), how they sequence across a stretch of discourse (§6.2), how context modulates which operations are avail-

able and how they are read (§6.3), why operations are legible primarily against the *anchor* of what they are responding to rather than in isolation (§6.4), and what the text-based formalization of Paper V structurally cannot see (§6.5). The methodological upshot for Paper III is specific: an annotation protocol that treats utterances as isolated semantic units, ignoring the composability, sequencing, and anchoring that this section describes, would miss a substantial fraction of what makes compensatory-defensive speech recognizable.

6.1 Composition within an utterance

The first observation is that operations are not mutually exclusive at the level of the single utterance. A compensatory-defensive utterance can, and frequently does, instantiate more than one operation simultaneously — and the patterns of this simultaneous instantiation are themselves structured, not arbitrary. We distinguish three principal modes of intra-utterance composition.

Within-family stacking. Operations from the same vector family co-occur with characteristic frequency because they share the same underlying precision configuration. A speaker in a DEFENCE-regime fight configuration is likely to combine, in a single utterance, a you-attribution with a global-quantifier deployment and an externalized-agency framing: *"****you always do this — you're trying to make me feel small, and now you've made me upset.****"* The three operations are technically distinct, but they issue from the same precision signature and reinforce one another; they are what the system produces when the same configuration is realized across the three positions where fight-family operations fit into a sentence (subject attribution, temporal quantifier, grammatical agency). Annotation must be able to mark all three rather than forcing a choice among them.

Cross-family stacking. Operations from different vector families can also co-occur, and when they do, the combination is usually not random: it reflects the speaker's attempt to perform more than one defensive function at once. A you-attribution delivered with hedge-stacking (*"****I don't know, maybe, I mean, you might kind of be trying to, like, make me feel bad?****"*) combines fight-family attribution with fawn-family self-minimization; the combination is recognizable as a specific sociolinguistic move, often produced by speakers whose protected prior includes a strong prohibition against straightforward aggression. An abstract-register shift that is simultaneously a meta-commentary (*"****it's interesting how conversations like this always seem to escalate****"*) combines flight-family distance with meta-positioning. These cross-family stackings are not exceptions to the alphabet; they are *compositions of the alphabet*, which is exactly what an alphabet is for.

Hierarchical embedding. A third, subtler mode is when one operation acts as the structural host for another. Consider: *"****I was just wondering — and please tell me if this is stupid — whether you've been avoiding me.****"* The outer structure is a fawn-family hedge-stack with apology surplus; embedded within it is a fight-family you-attribution (the accusation of avoidance). The operations are not merely simultaneous; they stand in a host-guest relationship, where the fawn shell makes the fight content deliverable without triggering the social consequences of bare fight expression. Hierarchical embeddings are diagnostic of certain speaker profiles — notably,

speakers whose protected priors include both "I must not be aggressive" and "this grievance must be communicated" — and their detection requires annotation that can represent nested, not merely parallel, operation labels.

The methodological consequence for Paper III is that annotation at the utterance level must allow *multi-label assignment with relational structure*: not just "this utterance instantiates operations A, B, and C" but, where relevant, "A hosts B, and C is parallel to both." The engine need not recover the full relational structure in every case; but the annotation ontology must be rich enough to represent it when the structure is clear, or important information is lost at the ingestion stage.

6.2 Sequential composition across utterances

The second observation concerns what happens across utterances rather than within them. Stretches of compensatory-defensive speech are rarely homogeneous; they have internal structure, and that structure is informative. Paper IV will develop the formalization of longer-scale trajectories and regime transitions; here we confine ourselves to the short-range patterns that are visible within a single conversational exchange of a few turns.

Three short-range patterns are worth naming, because they recur across speakers and contexts and because they constrain what an annotator (or engine) should expect to find.

Escalation chains. A speaker who begins an exchange with a mild you-attribution, meets interlocutor resistance, and responds with a stronger global-quantifier deployment is traversing an escalation chain within the fight family. The escalation is not random; it is the system attempting to restore the defensive efficacy of the initial operation when the interlocutor's response suggests the initial operation has failed to attenuate the predicted shame-signal. Escalation within a family is common; escalation across families is less common and usually indicates that the initial defensive configuration is failing structurally, prompting the system to try a different defensive mechanism.

Repair-failure sequences. A speaker in the COGNITIVE regime, facing a challenging topic, may deploy an ecological retreat ("let me think about this"), reconnect, and continue engaged. A speaker in the DEFENCE regime, facing the same topic, may deploy what *looks like* an ecological retreat (the abstract-register shift's surface is similar to a genuine request for reflection time) but then fails to reconnect: the topic is not resumed, the speaker's first-person position is not recovered, the affective register stays lifted. The diagnostic signal is not in the first utterance but in the failure of the expected repair. An annotator who marks only the first utterance misses the critical information; the critical information lives in the sequence.

Stability under challenge. A third pattern is the absence of expected movement. A speaker whose initial utterance instantiates, say, a hedge-stack, and who continues to produce hedge-stacked utterances across several turns despite interlocutor reassurance, is exhibiting a regime-level stability that a single utterance cannot disclose. The persistence of the operation-pattern across a challenge period is what distinguishes a regime commitment from a contextual response. This is the short-range precursor

of the regime-trajectory analysis that Paper IV will develop.

The consequence for annotation is that individual operations need to be labeled *within their sequential context* — at minimum, with the previous and next turns visible. An annotator looking at a single utterance has access to its operation-level composition (§6.1) but not to its regime-level significance; only a sequence of utterances begins to disclose regime. Paper III's engine reflects this: operation-level classification runs on the single utterance, but regime-level classification requires a context window of several turns.

6.3 Context as modulator

The third observation is that operations are not context-free. The same grammatical or pragmatic structure, considered in isolation, can be compensatory-defensive in one context and ecological in another — not because the operation is ambiguous at the level of linguistic form, but because the ecological/compensatory discriminator (§7) is itself context-sensitive. We distinguish three types of contextual information that modulate operation-reading.

Interlocutor context. The identity, status, and relational history of the interlocutor shape which operations are proportionate. A hedge-stack directed at a police officer during a traffic stop is proportionate to the power differential and the stakes; the same hedge-stack directed at a spouse in a casual dinner conversation is not. The operation is not different in its grammatical form; what differs is the fit between the operation's function (drain assertoric force, reduce exposure) and the situation's actual requirements. An annotation system that does not encode at least basic interlocutor information — role, relationship, power-differential — cannot make the ecological/compensatory distinction accurately.

Register context. The conventional expectations of the conversational register set a baseline against which operations are read. A formal work email that includes moderate hedging is operating within the conventional politeness expectations of its register; a casual message to a close friend that includes the same hedging is operating outside the conventional politeness expectations of *its* register, and the departure itself is informative. Register is, in this sense, the local normative prior against which operation-deployment is evaluated. The engine of Paper III uses register markers (vocabulary choice, syntactic formality, greeting and closing conventions) as input to the ecological/compensatory classifier for precisely this reason.

Immediate conversational prior. The most proximate contextual information is what the interlocutor just said. An operation deployed in response to a confrontational prompt is read differently from the same operation deployed in response to a warm prompt, for reasons we develop more fully in §6.4. The immediate prior is, in effect, the local evidence against which the operation's proportionality is evaluated. An engine that processes utterances independently, without looking at the interlocutor's prior turn, cannot perform this evaluation.

A fourth type of context, which we defer to later papers, is *speaker history*. The same utterance, produced by a speaker whose conversational history over the past month has been regime-stable in COGNITIVE, reads differently from the same utterance pro-

duced by a speaker whose history has shown persistent DEFENCE. Speaker-history context is what Paper IV and Paper V operationalize as trajectory-level information; at the operation level of Paper III, it is not directly available, but it constrains the prior distribution over regime classifications that the engine uses.

6.4 The anchor prompt

The single most important methodological point of this section — and the one that most distinguishes our annotation protocol from approaches that treat utterances as isolated semantic units — concerns what we will call the *anchor*: the specific prompt or utterance to which the target utterance is responding. Operations are legible primarily against their anchor, not in isolation. This is not a convenience; it is a structural property of the phenomenon.

Consider a hedge-stacked utterance: *"****I was kind of maybe just wondering if, I don't know, you might possibly have thought about it.****"* In isolation, this is recognizable as a fawn-family operation — the grammatical form unambiguously instantiates hedge-stacking. But the operation's *function*, and therefore its ecological-or-compensatory status, is determined by what it is responding to. As a response to *"****will you marry me?****"* the hedge-stack is doing serious work: the speaker is trying to express uncertainty about a life-altering commitment in a way that minimizes harm to the interlocutor. As a response to *"****do you want coffee?****"* the same hedge-stack is disproportionate, and its disproportion is diagnostic of a regime configuration in which even trivial interactions activate exposure-minimizing policies. The grammatical form is identical. The operation is identical. The interpretation is not.

This is the reason operations must be annotated *with their prompts*, not alone. An annotation corpus composed of isolated utterances, however carefully labeled at the operation level, loses the information that discriminates ecological from compensatory expression. A corpus that pairs each utterance with its immediate prior — and, ideally, with enough preceding context to disambiguate the register and the interlocutor relationship — preserves that information and makes the ecological/compensatory classification tractable.

The implication for Paper III's engine design is direct. The ingestion format is not "utterance plus label"; it is "prompt plus utterance plus label," with prompt including, at minimum, the immediately preceding interlocutor turn and, where available, a summary of the conversational stance established in the prior turns. The engine's operation-level classifier runs on the utterance; the engine's ecological/compensatory classifier runs on the *pair*. Paper III details how these two classifiers are coupled within the engine's layered architecture.

A practical note. In some source materials — transcribed therapy sessions, for instance — the anchor is immediately available because the structure of the source is explicitly dialogic. In others — memoir, first-person reflective writing, monologic discourse — the anchor is implicit and must be reconstructed from the speaker's own framing. The reconstruction introduces additional uncertainty, and the engine's confidence in its ecological/compensatory classification should be correspondingly lower

on monologic input than on dialogic input. This is one of the axes along which the engine reports calibrated uncertainty to the user.

6.5 Prosody and what text-based formalization misses

The final point is a limitation honestly stated. Our formalization operates on text, and the empirical work of Paper V is conducted on a text corpus. This is a principled choice — text is the observable with the broadest availability, the most transparent reproducibility, and the fewest confounds from transcription decisions — but it is a choice that systematically omits a significant channel of defensive-speech signal: prosody.

Spoken defensive speech carries prosodic signatures that text does not represent. Fight-family operations, when spoken, often show elevated volume, compressed pitch range, and accelerated pacing; fawn-family hedges show characteristic upward inflection at statement endings (the "uptalk" that converts declaratives into quasi-questions), elongated softening particles, and reduced volume on assertoric content; flight-family abstract-register shifts are often accompanied by a noticeable flattening of affective prosody, as the speaker distances not only semantically but phonetically from the topic; freeze-family epistemic vagueness shows characteristic prosodic flattening, reduced pitch variability, and extended pausing. These prosodic markers are real, they are detectable, and they are sometimes one of the most reliable channels for discriminating ecological from compensatory expression — a hedge delivered with warm, fluent prosody reads differently from a syntactically identical hedge delivered with flat, compressed prosody.

Text-based formalization structurally cannot access these markers. The consequence is twofold. First, our formalization is a *subset* of what a complete formalization of compensatory-defensive speech would contain; a future extension to multimodal data — audio, video, physiological — would enrich the alphabet and improve discrimination. Second, our formalization's accuracy on text should be understood as a *lower bound* on what the phenomenon is in principle accessible to: where text-based classification is uncertain, prosodic information would often resolve the uncertainty, and the field's long-term goal should include the development of multimodal annotation protocols that integrate the two.

For the present purposes, we note the limitation and proceed. Text is sufficient for the methodological claim Paper II is making — that compensatory-defensive speech admits finite-basis formalization — because the argument rests on the operations, which are linguistically realized and detectable in text. It is also sufficient for the empirical claim Paper V will make, which concerns the distribution of operations and regimes across a bilingual text corpus, not the full multimodal phenomenon. Multimodal extension is flagged as future work rather than as a present gap; the gap it closes is in scope, not in principle.

7. The ecological/compensatory discriminator at the operation level

Paper I §3.5 introduced three criteria for distinguishing ecological from compensatory expression of the biological defensive vectors: *proportionality* (does the response

match the magnitude of the present trigger), *reversibility* (does the response release when the situation resolves), and *contact-preservation* (does the response leave the interactive channel intact). These criteria were stated at the level of the overall defensive response — the vector as a whole, expressed across an interaction. For Paper III's engine to operationalize the distinction, we need the criteria to work at the level of individual operations, where they can be applied to specific speech acts rather than to diffuse behavioral patterns. This section shows how each criterion translates to the operation level, and shows that at this level each criterion is not a binary but a *continuous, context-sensitive variable* that must be evaluated against the contextual information §6 has specified. The three criteria are not independent; they covary in characteristic ways that together constitute the composite discriminator the engine deploys.

7.1 Proportionality

Proportionality at the operation level is the degree of fit between two magnitudes: the *magnitude of the operation* and the *magnitude of the trigger* to which it responds. Neither magnitude is a single number; both are multidimensional, and the fit between them is evaluated jointly, not component by component.

The magnitude of an operation has at least three components. First, *intensity*: a hedge-stack with six softeners is more intense than one with two, a you-attribution with a global-quantifier and a motive-attribution is more intense than a bare you-attribution. Second, *scope*: an operation that frames the entire interaction (global quantifier, externalized-agency framing across multiple turns) has larger scope than an operation confined to a single claim. Third, *commitment*: an operation delivered with high pragmatic force — flat assertoric register, first-person commitment to the defensive frame — is more heavily committed than one delivered tentatively.

The magnitude of a trigger has, correspondingly, several components. First, *stakes*: is the trigger about a life-altering commitment, a significant relational issue, a trivial logistical matter? Second, *novelty*: does the trigger introduce content the speaker has never encountered, content that challenges a well-established self-model, or content that is routine? Third, *power-differential*: does the trigger arrive from a superior, a peer, or someone with less social authority? Fourth, *confrontational intent*: is the trigger phrased as an attack, a neutral inquiry, or a warm overture?

Proportionality is the fit between these multidimensional magnitudes. A hedge-stacked utterance in response to "will you marry me?" is a high-magnitude operation in response to a high-magnitude trigger; the magnitudes fit, and the fit is what makes the operation ecological rather than compensatory, despite the operation's intrinsic strength. A hedge-stacked utterance in response to "do you want coffee?" is a high-magnitude operation in response to a low-magnitude trigger; the magnitudes do not fit, and it is precisely the *mis*fit that makes the operation compensatory. Proportionality is not about operation-strength per se; it is about the joint distribution of operation-strength and trigger-strength.

This is why proportionality cannot be reduced to a property of the utterance alone, as §6.4 already argued. The hedge-stack is the same utterance in both cases; its

proportionality differs because the trigger differs. Paper III's engine evaluates proportionality by estimating both magnitudes — from operation-level linguistic features for the operation, from anchor-prompt features for the trigger — and computing the fit as a joint measure. The output is not a binary proportionate/disproportionate label but a calibrated score reflecting degree of fit, with the engine reporting its confidence in each component estimate separately.

A caveat. The engine's estimate of trigger-magnitude is fallible precisely where it matters most: in ambiguous interpersonal exchanges where the speaker and the interlocutor may disagree about the stakes, the novelty, or the confrontational intent of the trigger itself. A speaker in the DEFENCE regime may experience a neutral question as a high-stakes challenge; the engine, looking at the text, sees the neutral question. This asymmetry — between the engine's view of the trigger and the speaker's view of the trigger — is itself diagnostic information. When a speaker's operation-magnitude is consistently high relative to the engine's estimate of trigger-magnitude, this is a signature of the speaker's DEFENCE configuration, not a failure of the engine's estimate. The engine reports the mismatch and lets the downstream analysis interpret it.

7.2 Reversibility

Reversibility at the operation level is a temporal property: it measures whether, and how quickly, an operation is released when the conditions that warranted it cease to obtain. Unlike proportionality, which can in principle be assessed at a single utterance with its anchor, reversibility is intrinsically a property of *sequences* — it requires observing what happens when the interlocutor signals that the defensive operation is no longer needed.

Consider two speakers, each of whom responds to a perceived slight with a you-attribution. The interlocutor, in both cases, acknowledges the slight and offers an apology. The ecological speaker, whose operation was proportionate to the trigger and is now no longer needed, releases the you-attribution: the next turn is no longer framed through attributive hostility, the relational channel is repaired, the defensive posture drops. The compensatory speaker, whose operation was not, at the deepest level, a response to this slight but a deployment of a chronically-configured fight-family policy, does not release: the next turn continues to attribute hostile internality, introduces new grievances, escalates. The interlocutor's acknowledgment — the thing that would, for an ecological speaker, terminate the defensive episode — does nothing to modify the compensatory speaker's operation-selection, because the operation was not, in fact, being selected in response to the acknowledgment-sensitive situation in the first place.

This temporal signature is what reversibility captures, and it is why reversibility cannot be evaluated on a single utterance. The engine needs a sequence — at minimum, the target operation, the interlocutor's response, and the speaker's subsequent turn — to classify reversibility. Longer sequences yield more reliable classification, because reversibility is a graded property: partial release, delayed release, and release-followed-by-recurrence are all distinct patterns that a sufficient window can discriminate but a short window cannot.

Three components structure the reversibility measurement. First, *safety signal detection*: did the interlocutor actually provide a signal that the defensive operation is no longer warranted — acknowledgment, apology, de-escalation, expression of concern? Without a safety signal, the question of reversibility does not arise; an operation that persists in the absence of a safety signal is not thereby irreversible, only unresolved. Second, *response latency*: if a safety signal was provided, how many turns pass before the operation is released? An ecological fight operation typically releases within one or two turns of the safety signal; a compensatory fight operation may persist indefinitely. Third, *completeness of release*: when release occurs, is it full (the operation no longer appears in subsequent turns), partial (the operation attenuates but persists), or merely surface-level (the grammatical form of the operation drops while the underlying precision configuration persists, manifested in other operations from the same family)?

The operationalization for Paper III requires the engine to track operation-persistence across a context window of typically 5–8 turns, with automatic detection of safety signals in interlocutor turns and computation of release patterns on speaker turns. The output is a calibrated reversibility score reflecting the degree to which the operation was released in response to the safety signal, with appropriate uncertainty when the safety signal is ambiguous or absent.

A noteworthy case. When reversibility is high but proportionality is low — a disproportionately-strong operation that nevertheless releases readily once the safety signal appears — the profile suggests an ecological-adjacent deployment: the speaker over-shot the magnitude but was genuinely responsive to repair. When reversibility is low but proportionality is high — a proportionate operation that nevertheless persists after the safety signal — the profile suggests a regime configuration in which even warranted defensive responses are held beyond their functional scope. The composite of the two criteria is informative in a way that neither is alone, a point we develop in §7.4.

7.3 Contact-preservation

Contact-preservation at the operation level measures whether the operation, while performing its defensive function, leaves the interactive channel intact for the interlocutor to respond within. This is the criterion that most sharply distinguishes named retreat from silent withdrawal, proportionate pushback from channel-damaging attack, and acknowledged not-knowing from dissociative evacuation.

Two operations with identical defensive functions can differ dramatically in their contact-preservation signature. Compare an ecological flight-family operation — *“****I need to step back from this conversation; I’ll come back to it tomorrow****”* — with a compensatory flight-family operation — an abstract-register shift that lifts the speaker out of the topic without naming the move. Both perform retreat. The first names the retreat, attributes it to the speaker (“I need”), and explicitly preserves the possibility of return (“tomorrow”). The second does none of these things: the retreat is executed without acknowledgment, the speaker’s agency is not foregrounded, the conditions for return are not specified. The interlocutor, faced with the first, knows what happened and can plan accordingly; the interlocutor, faced with the second,

receives a discontinuity without information about its source or duration.

The contact-preservation measurement has four principal components, each of which is a continuous linguistic variable.

Addressee-orientation: does the operation remain oriented toward the interlocutor as a present, responsive agent, or does it withdraw the interlocutor from the interactional frame? A you-attribution, despite being a fight-family operation, at least addresses the interlocutor; an abstract-register shift that moves to "people like us" or "one could say" addresses no one in particular. The second-person pronoun, the direct question, the acknowledgment of the interlocutor's previous turn — these are all addressee-orientation markers whose presence supports contact-preservation and whose absence attenuates it.

Invitation to respond: does the operation leave room for the interlocutor to take the next turn meaningfully, or does it structurally preclude response? An operation that ends in a global-quantifier absolute ("you always do this") precludes response by denying the possibility of contextual repair; an operation that ends in a specific, present-tense claim invites response because the claim is in principle testable against the current interaction. The presence or absence of response-invitation is partly syntactic (question markers, trailing specificity) and partly pragmatic (engagement with what the interlocutor has said vs. monologic elaboration).

First-person presence: does the speaker's own first-person stance remain available in the operation, or has it been evacuated? An ecological fight operation ("I feel that you're disregarding my concern") preserves the speaker's first-person contour; a compensatory fight operation with externalized-agency framing ("you made me feel disregarded") evacuates first-person agency. When first-person presence is evacuated, the interlocutor has no one to respond *to*; the speaker has, in effect, removed the coordinates at which response could be directed.

Repair-receptivity: does the operation's pragmatic form leave the speaker available to receive and integrate a repair move from the interlocutor? This component overlaps with reversibility but is not identical to it: reversibility is a post-hoc measure of whether repair was received; repair-receptivity is a concurrent measure of whether repair could, in principle, be received at the moment of the operation. An epistemic-vagueness operation that declines to commit to any propositional content leaves the speaker unavailable for repair, because there is no proposition the interlocutor can address.

The engine operationalizes contact-preservation as a weighted composite of these four components, each of which can be estimated from operation-level and discourse-level linguistic features. As with the other criteria, the output is a calibrated score, not a binary judgment, and the engine reports its confidence separately for each component. The composite score is not a simple average; some components are more diagnostic in some contexts than in others (addressee-orientation weighs more heavily in dyadic than in multi-party exchanges; first-person presence weighs more heavily in affective than in informational exchanges), and the engine's weights are context-sensitive accordingly.

7.4 Covariance and the composite discriminator

The three criteria are not independent. In practice, they covary in characteristic patterns that themselves carry diagnostic information — patterns that are predicted by the precision-regime architecture of §5 and that the engine uses to construct a composite discriminator more sensitive than any single criterion alone.

The DEFENCE regime characteristically expresses all three criteria in the compensatory direction: disproportionate operation-magnitude relative to trigger-magnitude (because the operation is driven by the protected prior rather than by the current trigger), low reversibility (because the protected prior does not update in response to interlocutor-provided safety signals), and low contact-preservation (because the policies cached under the protected prior were selected for internal attenuation of predicted shame, not for preservation of the interactional channel). The COGNITIVE regime characteristically expresses the opposite pattern: proportionality, reversibility, and contact-preservation together, with each criterion at levels appropriate to the current context. These are the two poles the composite discriminator is designed to distinguish, and when all three criteria align, the classification is straightforward.

The more interesting cases are the hybrids, where the three criteria disagree. A speaker may produce an operation that is proportionate to the trigger but has low reversibility: the initial magnitude fits, but once the operation is deployed it is not released even when the interlocutor signals that release would be appropriate. A speaker may produce an operation with high contact-preservation but low proportionality: the operation is disproportionately strong for the situation, but it is delivered in a way that preserves the channel for response. A speaker may produce an operation with high proportionality and high reversibility but low contact-preservation: the operation is appropriate in magnitude and is released when the situation resolves, but while it is active, it damages the channel.

Each of these hybrid profiles corresponds to a distinct speaker-configuration that the composite discriminator can, in principle, identify. Proportionate-but-irreversible operations are often diagnostic of ILLUSION-regime dynamics (§5.4): the initial deployment tracks the trigger reasonably, but the operation, once deployed, is held by the coherence-maintenance machinery of the ILLUSION configuration even after the external pressure resolves. Disproportionate-but-contact-preserving operations often indicate a speaker whose regime is primarily COGNITIVE but who has been momentarily destabilized by a high-salience trigger — the composite profile is ecological-leaning despite the proportionality failure. Proportionate-and-reversible-but-channel-damaging operations often indicate a speaker whose regime is primarily COGNITIVE but whose repertoire of ecological operations is limited in a specific domain (e.g., a speaker who handles intellectual disagreement cognitively but lacks ecological operations for emotional conflict, and so produces channel-damaging operations when emotional conflict arises). The composite discriminator's classification is more nuanced than any single criterion permits.

The engine's composite discriminator is, accordingly, not a weighted sum over the three criteria but a joint classifier over the three-dimensional space they span. The classifier's output is a distribution over a small number of profile types — ecological, compensatory-DEFENCE, compensatory-ILLUSION, and several hy-

brid profiles — with calibrated uncertainty reflecting how clearly the observed operation-in-context falls into each profile. This is what Paper III's engine delivers as the ecological/compensatory layer of its analysis, and it is the layer that most directly distinguishes Paper II's framework from approaches that treat the ecological/compensatory distinction as binary or that ground it in operation-level features alone.

7.5 What the discriminator delivers and what it does not

Two clarifications close the section, both of which guard against over-reading what the composite discriminator can do.

First, the discriminator operates at the level of operation-in-context, not at the level of the speaker. A single operation classified as compensatory does not license a claim about the speaker's regime; it licenses, at most, a claim about the precision-configuration active in that moment of speech. Regime classification requires aggregation across many operations, in the way §6.2 described for sequential composition and §5.6 described for operation-to-regime inference. A speaker may produce compensatory operations in some contexts and ecological operations in others; the discriminator reports what it sees operation by operation, and the regime-level claim is a downstream inference from the accumulating pattern. Paper III specifies the aggregation mechanism; Paper II does not.

Second, the discriminator is not an evaluative instrument. Classifying an operation as compensatory is not a judgment that the operation is wrong, or that the speaker is defensive in a morally-loaded sense, or that the speaker should deploy a different operation. Compensatory operations are ordinary human linguistic behavior; every speaker produces them; they are produced for reasons that, within the precision-regime architecture, are mechanically sensible — they attenuate predicted shame-error, which is what the system is configured to do. The discriminator identifies a property of the operation's deployment relative to its context; it does not prescribe. The prescriptive work — therapeutic, pedagogical, or self-reflective — is work that can be built *on top of* the discriminator's classification, but it is not work that the discriminator itself performs, and Paper II takes no position on what such prescriptive work should look like.

These two clarifications together establish the scope of the ecological/compensatory machinery introduced in this section. The machinery is a classification instrument operating at a specific level of description, producing calibrated outputs over a bounded set of profile types, for downstream use in analyses that remain to be specified elsewhere in the series. It is not, and is not claimed to be, a diagnostic tool, an evaluative metric, or a prescriptive guide.

8. Scope, three levels of description, and bridge to Paper III

8.1 What this paper does not claim

Six scope restrictions follow from what has been argued, each of which we state explicitly because the formalization is easier to misread than to read correctly.

First, we do not claim that speech in general is combinatorial in the sense developed here. The claim is restricted to the specific subclass of speech characterized in Paper I as compensatory-defensive — speech produced under predictive-shame dynamics with protected priors. Speech produced outside those dynamics — genuine inquiry, cooperative problem-solving, affectionate exchange, creative generation — lies outside the frame and requires different accounts. The combinatorial architecture we have specified is a claim about a subsystem, not about language.

Second, we do not claim that the operations listed in §4 are exhaustive. The list is provisional and will be refined through Paper III's operationalization. Some operations that we have grouped as variants of a single rule may prove, under detailed analysis, to be distinct; some operations listed as distinct may prove to be surface variants of a single underlying rule. The alphabet is not the last word; it is the first draft from which empirical work proceeds.

Third, we do not claim that the alphabet is language-universal. The empirical work reported in Paper V tests the alphabet on a bilingual English–Russian corpus, and we expect cross-linguistic variation in how operations are realized (specific linguistic markers, grammatical patterns, prosodic signatures). What we claim is invariance at the level of operation-identity: the operation that Russian speakers realize through a particular morphosyntactic pattern and English speakers realize through a different construction is the same operation in the sense relevant to our analysis. That claim is testable and will be tested.

Fourth, we do not claim that operation-identification is sufficient for clinical assessment. Clinical judgment remains necessary for the work clinical judgment does — integrating history, relational context, transferential dynamics, and treatment plan. The formalization's role is to support clinical judgment, to make visible to clinicians the operational patterns they may already sense but have no common language for, and to enable research at scale that clinical observation alone cannot support. It is not to replace clinical judgment.

Fifth, we do not claim that the engine of Paper III implements the full alphabet. Paper III begins with a constrained subset and scales; parts of the alphabet articulated in the present paper will not be operational in Paper III's first release, and some may not be operationalizable at all within the constraints of a current-generation classification architecture. Where the alphabet specifies more than the engine implements, the alphabet is the theoretical commitment and the engine the empirical approximation.

Sixth, we do not claim that all precision dynamics in discourse conform to four regimes. The four-regime architecture of §5 applies specifically to compensatory-defensive dynamics as characterized in Paper I. Discourse outside those dynamics — most of ordinary speech — traverses a much larger precision-configuration space; the four regimes are not a general taxonomy of discourse but a specific claim about the attractors that emerge under predictive-shame pressure.

A seventh commitment, methodological rather than scope-defining, is consistent with what Paper I §2.3 announced for the series as a whole. Each paper in this series includes explicit delineation of what is and is not claimed, and — where implementations are presented — honest analysis of where those implementations fall short of their intended scope is treated as part of the contribution. In Paper II this takes the

form of the six scope restrictions just listed. In Paper III it will take the form of a calibrated limitations section against code-verified audit. In Paper V it will take the form of detailed discussion of failure modes against gold-standard annotation. Readers oriented toward honest failure reporting as a methodological standard will recognise this consistent treatment across the series.

8.2 Three levels of description

The architecture we have developed plugs into a larger series, and that larger series invokes three terminological systems that risk being heard as three competing taxonomies. Before we proceed to the bridge with Paper III, we make the relation among them explicit.

The three systems are: the biological **vectors** introduced in Paper I, the **precision regimes** developed in §5 of the present paper, and the **engine-level regimes** — which Paper III refers to as attractors — named in Paper III. Each addresses a different question, and the reader who expects them to collapse into a single hierarchy will be correctly frustrated, because they do not collapse in that way.

Vectors (Paper I) describe *direction of motivation*. What is the organism trying to do: attack a proximate threat, withdraw from it, appease it, immobilize before it, or orient toward what is unknown? These are phylogenetically ancient behavioral-affective policies with well-documented neural substrates. Paper I develops five: fight, flight, fawn, freeze, and — added in Paper I §3.5 as the non-defensive orientation — SEEK. Vectors are pre-linguistic; they are what the organism inherits from its evolutionary history, and they survive in speech as directional logic.

Precision regimes (§5 of the present paper) describe *configuration of the inferential machinery*. Which priors are rigid, which evidence channels are attenuated, how much weight is placed on self-protective priors relative to disconfirmatory input, whether higher-order inference over precision itself is available. These are architectural states of the hierarchical generative model, not policies about what to do. The four regimes — DEFENCE, COGNITIVE, ILLUSION, META — are attractor configurations of this precision-allocation space.

Engine-level regimes (Paper III) describe *observable speech-level patterns* that emerge from the joint operation of the first two. The ten engine regimes enumerated in Paper III (BUILD, SEEK, UNSEAL, LOCK, SEAL, DRAIN, FLOOD, EDGE, SHIFT, VOID) are empirically-grounded clusters in the space of actual discourse: patterns recurring consistently enough across speakers and contexts that a classification engine can learn to recognise them.

The key conceptual point — and what resolves the apparent tension of "why five, why four, why ten" — is that the first two systems are *orthogonal*, not hierarchical. Vectors and precision regimes describe different things. A given vector can be expressed under any precision regime. Fight in COGNITIVE is the assertion of a proportionate boundary with attention to feedback and readiness for repair. Fight in DEFENCE is the accusatory you-attribution that does not release when the interlocutor acknowledges. Fight in ILLUSION is the persistent sense that the other is hostile even in the absence of any present trigger. These are the same vector — outward-directed mo-

bilized force — operating in three different configurations of the inferential machine. The configurations are not properties of the vector; they are properties of the system that is deploying the vector.

The same orthogonality applies to the other four vectors, which is why the product space of vector × regime has more cells than either system alone would have rows. Not every cell is equally populated in actual discourse — some vector-regime combinations are common, others rare, some perhaps empty — but the architectural space is defined by the Cartesian product, not by either system's internal distinctions.

Engine-level regimes, then, live in this product space. They are not redundant with either system; they are empirically-validated clusters of *which* cells of the vector × precision-regime space actually populate actual speech. An engine regime such as LOCK is approximately "fawn operations in DEFENCE configuration with high commitment and low reversibility" — a specific cell in the product space that recurs frequently enough to warrant its own label. An engine regime such as BUILD is approximately "fight and SEEK operations in COGNITIVE configuration with preserved META-accessibility" — a different cell in the same space. The ten engine regimes are not derived from vectors alone or from precision regimes alone; they are derived from the distribution of actually-occurring combinations.

This architecture has a consequence that matters for how the series is read. A reader who asks "but which vector is LOCK?" or "which precision regime is SEEK?" is asking a question the architecture cannot answer in isolation: LOCK is a vector-regime combination, and engine-level SEEK (to which we turn in §8.3) is a cluster that co-occurs preferentially with COGNITIVE precision allocation and biological SEEK-vector activation, but the cluster is the co-occurrence pattern, not the vector or the precision regime in isolation. The three systems map a single reality from three different angles; they do not stack.

8.3 A note on the word "SEEK"

The word *SEEK* appears in this series at two different levels of description, and since the same word is used at both, a reader who does not notice the distinction will conflate them. We name the distinction here.

Biological SEEK is the appetitive-motivational system introduced in Paper I §3.5 — the Panksepp (1998) SEEKING system, neurologically instantiated in the mesolimbic dopaminergic pathway (ventral tegmental area, nucleus accumbens, lateral hypothalamus, and cortical projections), functionally organising approach-toward-appetite, exploratory behaviour, and anticipatory engagement. Its formal counterpart in the predictive-processing framework is action for epistemic value — action selected to reduce uncertainty about states of the world rather than to secure preferred outcomes (Friston et al., 2015). Biological SEEK is a vector, in the sense developed in §8.2: a direction of motivation, a class of policies the organism can deploy.

Engine SEEK is one of the ten engine-level regimes enumerated in Paper III. It denotes a specific speech-level pattern: genuine inquiry moves — open questions, requests for elaboration, information-seeking turns — without concurrent push toward action or commitment. An engine SEEK attribution by the Paper III classifier means

that the speaker's current discourse configuration is predominantly this pattern; it does not, on its own, tell you anything about the speaker's underlying neural SEEKING system or the precision regime under which their generative model is currently configured.

The two are related but distinct. The engine regime SEEK is approximately the characteristic speech-level signature that biological SEEK produces when it is expressed ecologically — under COGNITIVE precision allocation, without co-activation of defensive vectors, with preserved accessibility to higher-order inference. That is why they share a name: the engine regime is the observable face of the vector in its healthy deployment. But the engine regime is not the vector. A speaker whose biological SEEKING system is dysregulated (suppressed in dorsal-vagal collapse, or chronically elevated in addiction) can nonetheless produce speech that, in surface form, reads as engine SEEK. Conversely, biological SEEK-vector activation can express itself, under compensatory conditions, in speech patterns that the engine would classify as something other than SEEK — as rumination, as inquiry-as-avoidance, as compulsive meaning-seeking. The mapping from vector to engine regime is one-to-many in ecological expression and becomes more complex under compensatory conditions.

When the present paper speaks of SEEK without modifier, in §§1–7, it means the biological vector as introduced in Paper I. When Paper III speaks of SEEK as one of its ten regimes, it means the observable speech-level pattern. The reader proceeding through the series should hold both senses and let context resolve the reference, as with any other term that operates at multiple levels of description in a mature scientific literature.

8.4 A structural note on META

A brief flag about META, the fourth precision regime of §5. The present paper has treated META as one of four regimes on parallel footing — DEFENCE, COGNITIVE, ILLUSION, META — because this presentation made the four-way contrast easier to develop and because it echoes the four-vector organisation that characterises the rest of the architecture. But there is an alternative reading that we want to acknowledge, because it has genuine force and because a reader who arrives at it independently may reasonably wonder why we did not.

On the alternative reading, META is not a fourth configuration of precision allocation in the same sense as the other three; it is a second-order operation that can modify any of them. DEFENCE + META is a speaker in defensive configuration who is nonetheless able to observe the fact of their defensiveness. COGNITIVE + META is ordinary reflective thought. ILLUSION + META is the phenomenologically striking state of recognising one's illusion while still being inside it. On this reading, META is best characterised as higher-order inference *over* precision, an operator rather than a state, and the architecture is really three first-order regimes with an orthogonal modifier rather than four parallel regimes.

We do not resolve this tension here. The four-regime presentation has the advantages of simplicity and formal parallelism and is adequate for the work the rest of the series does with it. The second-order reading has the advantages of phenomenological fidelity — META does feel like a perspective over the others rather than a sibling of

them — and of better tracking what Paper III's engine is doing when it makes META-sensitive classification decisions. We flag the tension as a structural question that future theoretical work in the series may revisit, and we proceed with the four-regime formulation in what follows because that is what §5 formally developed.

8.5 Bridge to Paper III

With the three levels of description made explicit and the two terminological flags in place, we can now specify cleanly what Paper III takes from this paper and what it adds.

Paper III takes three things as input. *First*, the alphabet of §4 — the vector-organised families of speech-level operations — serves as the provisional hypothesis space for Paper III's annotation ontology. *Second*, the four precision regimes of §5 serve as layers of analysis: classificatory dimensions along which Paper III's engine evaluates an utterance or sequence, not as labels to be assigned but as architectural coordinates within which label assignment is performed. *Third*, the operation-level ecological/compensatory discriminator of §7 serves as the engine's calibration target: the criteria by which operations receive continuous — not binary — scores for proportionality, reversibility, and contact-preservation, combined into the composite discriminator that Paper III operationalises as the hold_capacity scalar in the interval [0, 1].

Paper III adds what the present paper does not and could not provide. *First*, it specifies the annotation protocol with concrete linguistic features, inter-rater reliability procedures, and the operational cuts by which each operation of the alphabet is identified in real text. *Second*, it presents the Mindloom engine architecture as *structure-constrained LLM inference* — a classification pipeline in which rule-based layers (keyword detection, small fine-tuned classifiers, ontological constraints) define a bounded hypothesis space within which a large language model exercises classification judgment. This is the architectural contribution that operationalises what the present paper has argued is formally admissible. *Third*, Paper III introduces the ten engine-level regimes (BUILD, SEEK, UNSEAL, LOCK, SEAL, DRAIN, FLOOD, EDGE, SHIFT, VOID) as the empirically-validated clusters in the vector × precision-regime product space, derived from pilot annotation and stable across the calibration corpus. *Fourth*, it provides the limitations and honest-failure analysis that the methodological commitment of §8.1 foreshadows: a code-verified audit of where the engine's performance agrees with theoretical expectation and where it does not.

The division of labour between the two papers is thus: Paper II establishes the *warrant* for ontology-bounded classification of compensatory-defensive speech — it argues that the hypothesis space within which such classification operates is finite, structured, and derivable from independent theoretical considerations. Paper III provides the *instrument* — the specific engine that operates within that warranted space. The warrant does not automatically certify the instrument; Paper III's empirical limitations are its own to state. But the instrument does not make sense without the warrant; a classification pipeline without a principled hypothesis space is just a feature detector, and that is not what Paper III claims to be.

A final note on division of labour with Paper IV. Where Paper III classifies static configurations (what regime, what operations is this utterance or sequence manifesting),

Paper IV treats temporal dynamics (how regimes migrate across turns, how trajectories in the attractor space can be predicted, what developmental signatures emerge across extended dialogue). The four precision regimes of §5 are static attractors in this paper; in Paper IV they become basins of a dynamical system whose trajectories have their own formal characterisation. The present paper therefore does not address trajectory, transition, or prediction in any detail; those belong to the next paper in the sequence.

9. Conclusion

The question with which this paper opened was methodological: is the architecture of compensatory-defensive speech described in Paper I formalizable in a way that makes the speech, at least partially, computable? We have argued that it is, and we have done so along two coupled axes.

The first axis is a finite operational basis. Compensatory-defensive speech is generated from a bounded alphabet of speech-level operations — not templates, not fixed phrases — organised by the four defensive vectors of Paper I and composed according to context-sensitive rules. Four convergent arguments — metabolic, clinical, predictive-computational, and developmental — establish that this basis is finite as a matter of architectural mechanism, not as a matter of empirical accident. The sequences the alphabet generates are unbounded; the alphabet itself is not.

The second axis is precision regimes as emergent attractors. The same operations, viewed as temporal patterns of precision allocation, cluster into four recurrent configurations of the hierarchical generative model: DEFENCE, COGNITIVE, ILLUSION, META. These are not categories we imposed on the data; they are attractor-like configurations of precision allocation, emerging from the predictive-processing framework as solutions to the classes of predictive-shame situations that compensatory-defensive speech is deployed to manage.

Together, the two axes establish what the paper set out to establish: the hypothesis space for compensatory-defensive speech is both finite and theoretically principled. The hypothesis space is finite because the operational alphabet is bounded; it is principled because the boundedness follows from the predictive-processing framework rather than from engineering convenience, enumerative convenience, or any other external consideration. This is not a claim about what any particular engine can or cannot classify. It is a claim about the space within which classification, done in any way, would operate — and about why that space has the shape it does.

The methodological consequence is the warrant Paper III needs. An ontology-bounded classification architecture for compensatory-defensive speech is not, on this account, an engineering choice among alternatives; it is the formalisation the underlying architecture specifically authorises. Paper III operationalises this warrant as a concrete engine — structure-constrained LLM inference, with calibration, with honest limits-reporting, and with the empirical work that a warrant alone cannot do. The warrant and the instrument are distinct contributions, and neither supersedes the other.

What this enables for the field is, we think, worth stating without overstatement.

The study of defensive speech has been limited for decades by the absence of a shared operational language linking clinical observation to computational work. Clinicians have named the phenomena; computational approaches to related problems have largely operated through surface-feature detection; neither tradition has had access to what the other needs. The architecture articulated across Papers I–III offers a shared operational language — vectors from biology, operations from linguistics, regimes from predictive neuroscience, discriminators from clinical judgment — within which the two traditions can finally address the same object with compatible tools. Papers IV, V, and VI will develop the empirical and applied consequences. The present paper does one thing, and we hope it does it cleanly: it specifies the formal space within which that development can proceed.

Appendix A — Formal characterization

The four precision regimes developed in §5 are given a formal characterization below. The purpose is to define the hyperparameter space within which regimes are distinguished, to exhibit a minimal worked numerical illustration, to extend the formalism to dyadic interaction, and to list the five falsifiable predictions that the formalism generates. The mathematical material is adapted from prior theoretical work in the same programme; what the present appendix contributes is the placement of that material in the computability-warrant architecture of the main text.

Четыре precision regime, развитые в §5, получают формальную характеризацию ниже. Цель — определить пространство гиперпараметров, внутри которого регимы различаются, предъявить минимальную проработанную численную иллюстрацию, расширить формализм до диадического взаимодействия и перечислить пять фальсифицируемых предсказаний, которые порождает формализм. Математический материал адаптирован из предшествующей теоретической работы в той же программе; собственный вклад настоящего аппендикса состоит в размещении этого материала в архитектуре computability-санкции основного текста.

A.1 Variational free energy and precision-weighted message passing

Under the variational formulation of active inference (Friston, 2010), an agent maintains a generative model $p(o, s) = p(o|s)\, p(s)$ relating observations o to latent states s, and minimises variational free energy $F = E_q[\ln q(s) - \ln p(o, s)]$, which upper-bounds surprisal $-\ln p(o)$. In predictive-coding implementations, belief updates take the form of gradient descent on F with respect to posterior means μ (Friston, 2005).

In linear-Gaussian settings, precision is explicitly inverse variance. If the likelihood is $p(o|s) = N(s, \Pi_o^{-1})$ and the prior is $p(s) = N(\mu_0, \Pi_s^{-1})$, the posterior mean satisfies

$$\mu_{post} = (\Pi_s \cdot \mu_0 + \Pi_o \cdot o) / (\Pi_s + \Pi_o),$$

where $\kappa = \Pi_o / (\Pi_s + \Pi_o)$ is the Kalman gain — the weight given to incoming evidence relative to prior expectation. This equation is the formal object used throughout the appendix. It makes concrete how precision regimes differ: regime shifts are shifts in Π_s, Π_o, or both.

Extensions to hierarchical settings replace μ_0 with message-passing from higher levels, introduce precision-weighted prediction errors ε at each level, and extend to policy selection via expected free energy G. These extensions do not change the basic relation between precision allocation and inference trajectory that the rest of the appendix exploits.

A.2 Precision regimes as hyperparameter sets

We define a precision regime as a structured set of hyperparameters

R = { Π_o(l, m), Π_s(l, m), Π_π(l), γ(l) },

where l indexes hierarchical level, m indexes modality (auditory, semantic, interoceptive), Π_o are likelihood precisions, Π_s are prior precisions, Π_π are policy precisions, and γ denotes the relative weight of epistemic value in expected free energy. The four regimes introduced in §5 are specified as constraints on this hyperparameter space:

- **DEFENCE**: elevated Π_s on self-protective priors in threat-relevant subspaces; reduced Π_o on disconfirmatory channels; reduced γ (epistemic actions risk destabilising the protected prior).
- **COGNITIVE**: moderate, context-sensitive Π_s and Π_o; γ preserved, supporting information-seeking and cooperative repair; prior means calibrated.
- **ILLUSION**: biased prior means μ_0 in specific subspaces, and/or elevated Π_s with selective reduction of Π_o on channels that would disambiguate the bias; policy selection avoids disambiguating evidence (low effective γ in relevant subspaces); self-sustaining because confirmatory evidence is over-weighted and disconfirmatory evidence under-weighted.
- **META**: higher-order inference over Π_o and Π_s themselves, with learning of volatility and meta-precision enabling flexible transitions between the three first-order regimes; epistemic actions become explicit choices.

These specifications are theoretical constraints, not fitted parameterisations. Their purpose is to define a space of computational models that can be distinguished by empirical work, not to commit to specific numerical values that any given speaker is claimed to exhibit.

A.3 Toy model: a minimal two-layer linear-Gaussian regime switcher

A.3.1 The model We provide a minimal worked example that captures DEFENCE as prior-dominated updating, ILLUSION as biased priors plus reduced sensory precision, and META as inference over precision.

Let s be a latent discourse-relevant state and o an observation.

- Likelihood: $p(o|s) = N(s, \Pi_o^{-1})$.
- Prior: $p(s) = N(\mu_0, \Pi_s^{-1})$.
- Posterior mean: $\mu_{post} = (\Pi_s \cdot \mu_0 + \Pi_o \cdot o) / (\Pi_s + \Pi_o)$.

Regime instantiations (illustrative values):

- **COGNITIVE**: $\Pi_s = \Pi_o = 1$ (balanced).
- **DEFENCE**: $\Pi_s = 10$, $\Pi_o = 1$ (prior dominates).
- **ILLUSION**: biased $\mu_0 \neq s^*$ (target state) and reduced Π_o (e.g., $\Pi_o = 0.3$).
- **META**: $\Pi_o = \exp(\lambda)$, with λ inferred from prediction-error statistics.

A.3.2 Worked numerical illustration Assume $\mu_0 = 1$ (prior belief) and $o = 0$ (strong counterevidence). Compare regimes:

- **COGNITIVE**: $\mu_{post} = (1 \cdot 1 + 1 \cdot 0) / (1 + 1) = 0.50$ — balanced integration; belief moves halfway toward evidence.
- **DEFENCE**: $\mu_{post} = (10 \cdot 1 + 1 \cdot 0) / (10 + 1) \approx 0.91$ — belief barely updates despite contradictory evidence; Π_s dominates.
- **ILLUSION** ($\Pi_o = 0.3$): $\mu_{post} = (1 \cdot 1 + 0.3 \cdot 0) / (1 + 0.3) \approx 0.77$ — partial update that remains far from evidence; sensory precision is suppressed.

Two observations matter. *First*, psychologically similar "resistance to evidence" can arise from distinct computational mechanisms — either by increasing Π_s (DEFENCE) or by decreasing Π_o (ILLUSION). The same behavioural signature (μ_{post} close to μ_0, far from o) can have two structurally different causes. *Second*, DEFENCE and ILLUSION are behaviourally similar at the level of posterior mean but differ sharply in *temporal signature*: DEFENCE is threat-triggered and resolves when threat cues are removed; ILLUSION is self-sustaining and persists after threat removal. The temporal dissociation (elaborated as Prediction 2 in §A.5) is the key empirical handle for distinguishing the two regimes.

This two-parameter illustration does not exhaust the model's behaviour — the full model includes hierarchical structure, policy selection, and modality-specific precisions. The illustration is intended to make concrete what the hyperparameter specifications of §A.2 deliver computationally, not to substitute for the full formalism.

A.3.3 Simulation pseudocode The following pseudocode illustrates the qualitative coupling between precision and updating across regimes. It is intentionally abstract; the theoretical object is the coupling structure, not a specific implementation.

Algorithm 1. Precision-regime speech inference.

Inputs: μ_0, Π_s, λ_0, η (meta learning rate), regime_schedule[t], data[t]

For each t: regime ← regime_schedule[t]

if regime ∈ {DEFENCE, COGNITIVE, ILLUSION}: $\Pi_o \leftarrow \Pi_{o\_fixed}$[regime] else: // META $\Pi_o \leftarrow \exp(\lambda)$

$\mu \leftarrow (\Pi_s \cdot \mu_0 + \Pi_o \cdot o_t) / (\Pi_s + \Pi_o)$ $\varepsilon \leftarrow o_t - \mu$

if regime == META: $\lambda \leftarrow \lambda + \eta \cdot (\varepsilon^2 - \text{target})$

Record: μ, ε, Π_o

A.3.4 Expected simulation signatures Under a regime schedule such as COGNITIVE → DEFENCE → META, the model predicts three qualitative signatures:

- The posterior mean μ is "pinned" during DEFENCE (minimal movement toward evidence), relaxes back toward evidence under META as precision is recalibrated, and tracks evidence proportionately under COGNITIVE.
- Prediction error ε accumulates during rigid regimes (where μ cannot move to match o) and reduces after precision recalibration in META.
- Inferred precision Π_o during META shows adaptive modulation driven by error statistics — reduced when the environment is volatile, increased when it is stable — with the learning rate η controlling the timescale.

These are the signatures that a time-series analysis of regime-schedule simulations should exhibit, and they are what the empirical paradigms of §A.5 are designed to probe.

A.4 Extension to coupled dyadic inference

A.4.1 Regime-dependent coupling Communication inherently involves two agents; the single-agent formalism above must be extended. Following Friston and Frith (2015), dialogue is modelled as a coupled dynamical system in which each agent's speech acts become the other's observations.

Let Agent A be in regime R_A and Agent B in regime R_B. Each agent's posterior is updated according to its own precision settings:

μ_A(t+1) = f(μ_A(t), a_B(t); Π_A(R_A)), μ_B(t+1) = f(μ_B(t), a_A(t); Π_B(R_B)),

where a_X(t) denotes Agent X's speech acts at time t (determined by policy selection under expected free energy) and Π_X(R_X) is the precision configuration determined by Agent X's current regime.

A.4.2 Regime pairing dynamics Different regime pairings yield qualitatively different interaction dynamics. Four pairings are particularly diagnostic.

- **COGNITIVE–COGNITIVE**: mutual evidence-sensitivity supports convergent belief updating; both agents' prediction errors drive alignment, approximating the cooperative baseline of standard interactive-alignment accounts.
- **DEFENCE–COGNITIVE**: asymmetric coupling. The DEFENCE agent's reduced Π_o on disconfirmatory channels attenuates the influence of the COGNITIVE agent's signals; alignment is limited. The COGNITIVE agent may experience accumulating prediction error that ordinary conversational moves cannot resolve, potentially triggering their own regime shift — into DEFENCE (reciprocation), or into META (recognition that the channel is blocked).
- **ILLUSION–ILLUSION**: if both agents share compatible narrative biases, the dyad can stabilise in a mutually reinforcing local minimum. Surface alignment may appear high (each agent's utterances confirm shared priors) while predic-

tive accuracy in novel contexts degrades. This is the dynamics that characterise closed ideological pairings.

- **META–COGNITIVE**: the META agent's higher-order inference enables adaptive precision adjustment in response to the COGNITIVE agent's signals. Repair is accessible, perturbations are handled gracefully, and this pairing is predicted to yield the most robust alignment and repair capacity. In clinical settings the therapist is ideally in META; in educational settings the teacher is ideally in META relative to students in COGNITIVE.

These qualitative predictions are what the empirical programme of §A.5 is designed to test.

A.5 Five falsifiable predictions

The framework above yields five predictions, each stated in falsifiable form with paradigm, measures, and competing hypothesis specified.

A.5.1 Prediction 1: Regime-conditioned alignment dissociates from generic alignment *Claim.* Conversational alignment is higher and more stable when both interlocutors are in COGNITIVE or META regimes; apparent alignment under ILLUSION dissociates — surface alignment may be preserved while predictive accuracy in novel contexts degrades.

Paradigm. Cooperative dialogue task with periodic perturbations (unexpected topic shifts; contradictory information).

Measures. Behavioural alignment indices; next-utterance predictability; repair frequency.

Competing hypothesis. Generic interactive alignment predicts that alignment correlates monotonically with priming and shared representations, without regime-dependent precision control.

A.5.2 Prediction 2: Precision proxies track regime transitions with regime-specific temporal signatures *Claim.* Transitions into DEFENCE correspond to measurable reductions in effective sensory precision that *resolve when threat cues are removed*. Transitions into ILLUSION show reductions in sensory precision that *persist after threat removal*. This temporal dissociation is the key empirical handle for distinguishing the two regimes, which are otherwise similar at the level of posterior-mean behaviour.

Paradigm. Threat manipulation (social evaluation; time pressure) followed by a safe recovery period, during conversation.

Measures. Pupil dilation (arousal/uncertainty proxy); gaze and eye-tracking (attention allocation); ERP markers of expectancy violation.

Competing hypothesis. A purely arousal-based account predicts monotonic stress effects without the specific pattern of regime-dependent recovery dynamics.

A.5.3 Prediction 3: Neural synchrony reflects regime-conditioned coupling *Claim.* Inter-brain synchrony during real interaction depends on regime pairing: META–COGNITIVE dyads show robust coupling in higher-order networks; DEFENCE–COGNITIVE dyads show reduced coupling and increased asymmetry.

Paradigm. Hyperscanning EEG/MEG or fMRI during goal-directed conversation.

Measures. Phase-based synchrony (theta/alpha/beta); cross-brain Granger causality or transfer entropy; hyperscanning effective connectivity.

Competing hypothesis. Stimulus-driven coupling accounts predict coupling primarily at sensory levels; the regime hypothesis predicts coupling differences even under matched stimulus structure.

A.5.4 Prediction 4: Neuromodulatory perturbations shift regimes via precision mechanisms *Claim.* Manipulations affecting neuromodulatory systems linked to uncertainty and precision (cholinergic and noradrenergic influences; Yu & Dayan, 2005) shift the balance between DEFENCE-like and COGNITIVE/META-like updating, measurable in discourse repair and expectancy-violation processing.

Paradigm. Pharmacological challenge (where ethically appropriate) or task-based manipulation of expected versus unexpected uncertainty.

Measures. Behavioural evidence integration; EEG markers; pupil responses.

Competing hypothesis. Non-precision accounts predict global performance changes without specific changes in evidence-weighting signatures.

This prediction is framed conservatively: it does not invoke specific pharmacological agents or speculative models of psychedelic action, but rather the well-established link between neuromodulatory systems and precision estimation (Yu & Dayan, 2005; Feldman & Friston, 2010).

A.5.5 Prediction 5: Metacognitive training amplifies META signatures *Claim.* Brief training in epistemic humility and explicit uncertainty reporting increases META-like inference and improves dyadic prediction accuracy under perturbation, relative to generic conversational training.

Paradigm. Randomised intervention: metacognitive training versus control, then conversation under ambiguity.

Measures. Confidence reports; calibration; repair frequency; prediction accuracy; neural coupling measures.

Competing hypothesis. Generic skill-learning explains improvements without higher-order precision modelling; Bayesian model comparison should favour meta-precision models if the theory is correct.

A.6 Analysis pipelines and summary

Across paradigms, we recommend analyses that map onto competing generative models: hierarchical Bayesian model comparison (testing prior-precision versus

likelihood-precision versus mean-bias accounts); dynamic causal modelling for neuroimaging (testing gain changes predicted by precision shifts); time–frequency analysis for EEG/MEG (testing oscillatory coupling shifts with regime); and metacognitive metrics such as meta-d′ (operationalising inference over inference; Maniscalco & Lau, 2012).

Table A.1 summarises the mapping of the four regimes to canonical active-inference components.

Acknowledgments

The structural development of this paper and the broader series *Defense, Self, and Speech: From Biological Vectors to Linguistic Form* was conducted in extended dialogue with Claude (Anthropic), who contributed to the articulation of the combinatorial argument, the four-line evidential convergence, the alphabet of speech-level operations, and the four-regime precision architecture. The author retains full responsibility for the content, claims, and conclusions of this work, and for any errors that remain.

References

[To be compiled and merged with Paper I references for consistency across series]

Core references specific to Paper II:

- Bion, W. R. (1962). *Learning from Experience*. Heinemann.
- Bowen, M. (1978). *Family Therapy in Clinical Practice*. Jason Aronson.
- Bowlby, J. (1980). *Attachment and Loss, Vol. 3: Loss*. Basic Books.
- Chomsky, N. (1957). *Syntactic Structures*. Mouton.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3).
- Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin, 115*(1).
- Dodge, K. A. (1980). Social cognition and children's aggressive behavior. *Child Development, 51*(1).
- Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215.
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443.
- Fonagy, P., Gergely, G., Jurist, E., & Target, M. (2002). *Affect Regulation, Mentalization, and the Development of the Self*. Other Press.

- Freud, A. (1936/1966). *The Ego and the Mechanisms of Defense*. International Universities Press.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2).
- Friston, K. (2019). A free energy principle for a particular physics. *arXiv:1906.10184*.
- Friston, K., & Frith, C. (2015). Active inference, communication and hermeneutics. *Cortex, 68*.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neural Computation, 29*(1).
- Hazan, C., & Shaver, P. R. (1987). Romantic love conceptualized as an attachment process. *Journal of Personality and Social Psychology, 52*(3).
- Hesse, E. (2016). The Adult Attachment Interview: Protocol, method of analysis, and selected empirical studies. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of Attachment* (3rd ed.). Guilford.
- Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.
- Kernberg, O. (1975). *Borderline Conditions and Pathological Narcissism*. Jason Aronson.
- Klein, M. (1946). Notes on some schizoid mechanisms. *International Journal of Psycho-Analysis, 27*.
- Main, M., & Goldwyn, R. (1998). *Adult Attachment Scoring and Classification System*. Unpublished manuscript, University of California at Berkeley.
- Main, M., & Hesse, E. (1990). Parents' unresolved traumatic experiences are related to infant disorganized attachment status. In M. T. Greenberg, D. Cicchetti, & E. M. Cummings (Eds.), *Attachment in the Preschool Years*. University of Chicago Press.
- McEwen, B. S. (1998). Stress, adaptation, and disease: Allostasis and allostatic load. *Annals of the New York Academy of Sciences, 840*.
- Mikulincer, M., & Shaver, P. R. (2016). *Attachment in Adulthood: Structure, Dynamics, and Change* (2nd ed.). Guilford.
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.
- Porges, S. W. (2011). *The Polyvagal Theory*. W. W. Norton.
- Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*.
- Schauer, M., & Elbert, T. (2010). Dissociation following traumatic stress. *Zeitschrift für Psychologie, 218*(2).
- Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Dutton.

- Sifneos, P. E. (1973). The prevalence of “alexithymic” characteristics in psychosomatic patients. *Psychotherapy and Psychosomatics, 22*(2–6).
- Taylor, G. J., Bagby, R. M., & Parker, J. D. A. (1997). *Disorders of Affect Regulation*. Cambridge University Press.
- Van der Kolk, B. (2014). *The Body Keeps the Score*. Viking.
- Walker, P. (2013). *Complex PTSD: From Surviving to Thriving*.
- Winnicott, D. W. (1960). Ego distortion in terms of true and false self. In *The Maturational Processes and the Facilitating Environment*.