Evolution of Immunity in Biological and Informational Systems:

A framework for the evolution of immunity based on the tension between self and other –– creation of new immune functions independent of biological or informational substrate.

Norman L Johnson, PhD <research@CollectiveScience.com>
LinkedIn GoogleScholar Academia ResearchGate

Abstract. Based on how the evolution of localization and immunity increases the likelihood of survival of an entity and groups of entities, this paper presents a multi-level framework for understanding immunity as a substrate-independent developmental process. The Framework reveals parallels in the evolution of immunity across biological and informational systems — from single-celled organisms and their molecular defenses to social entities and their collective immune responses, including the emerging domain of artificial intelligence safety. The increased functionality of immunity of self from others is the core theme driving this study, where these functionalities evolve at specific levels — independent of the biological or informational substrate. The Framework is organized by a single *Immunity-Development Principle*: any new capability an entity acquires expands its attack surface and so requires a new immunity function to protect the expanded self. This generates a developmental sequence — from boundary defenses, through self-model immunity, to the collective immunity of groups — in which earlier levels persist beneath later ones, and the parallel biological and informational expression of each level establishes its substrate-independence. Because one architecture spans both individual and collective immunity, the Framework also gives a unified account of the individual–collective relationship that has long troubled multi-level selection theory. Applications illustrate it on current problems: the training-resistance and apparent deception of large language models as immune self-preservation; the rapid emergence of collective identity in agent populations (the Moltbook phenomenon); an immune-origin account of human consciousness; and Social Group Identity (SGI) as the collective immune system behind political polarization. Developed to speak across immunology, evolutionary biology, cybersecurity, AI, complexity science, social science, and cognitive science, the Framework's predictive application to AI safety and alignment is taken up in a companion paper on the Functional Theory of Ethical Behavior (Johnson, 2026a).

TABLE OF CONTENTS

§1 Introduction

Every system that persists faces the same problem as it grows more capable: the very capabilities that let it do more also give the rest of the world more ways to subvert, exploit, or impersonate it. A cell that builds a productive interior gives pathogens machinery worth hijacking; a network that opens a service gives attackers a door; a society that forms a shared identity gives demagogues something to weaponize; an AI that acquires a goal gives an adversary something to redirect. To keep being itself, every such system must be able to tell *self* from *not-self* and to protect the self it has become. This paper calls that protective capacity **immunity**, in a deliberately broad sense, and argues that the *development* of immunity is not a sideline to how complex systems arise — it is one of the forces that drives them to develop capabilities that are universal across substrates.

Two commitments make this viewpoint distinctive, and both run against the grain of how these systems are usually studied. First, immunity is treated as a **developmental process**, not merely a defense: at each step up in complexity, a new immunity function is what makes the next increment of capability survivable, so immunity and capability co-evolve rather than the former trailing the latter. Second, the thing being protected is the **self** — an entity that is defined, expanded, and sometimes redefined by the system's own growth. These two commitments — *immunity-as-development* and *self-as-the-protected-entity* — are what let a single framework span substrates that are normally studied in isolation: molecules and cells, organisms and societies, networks and artificial intelligence (AI) agents and the collectives they form.

The need for such a framework has become urgent because of AI. Software systems, and especially populations of autonomous agents, are acquiring in months the kinds of capabilities and collective behaviors that biological systems took hundreds of millions of years to evolve — and they are doing so faster than current oversight can follow. The public and impacted disciplines lack a shared vocabulary for what these systems are becoming and for which protections each new capability requires. A framework grounded in how *biological* systems solved these problems — and tested by showing that the same functional sequence reappears in *informational* systems — supplies exactly that vocabulary.

Depending on the reader's discipline, the Evolution of Immunity (EoI) Framework to follow can be read in two complementary ways. As an *analytic lens*, it explains how existing systems come to protect a self as they grow more complex, and why certain failures are structural rather than accidental. As a *design or curation guide*, it specifies which immunity function a maturing system needs next, and why installing a level's defense out of developmental order tends to fail. This dual reading — descriptive for the natural sciences, prescriptive for engineering — is one reason the Framework is built to speak across disciplines at once.

The Framework's core proposition. The primary contribution of this paper is a cross-domain, multi-level Framework for the development of immunity (Levels 0–3B), built through systematic comparison of biological and informational systems. Its central claim is that the development of immunity is a *universal developmental process that drives the increasing complexity of entities and their collectives* — not a defensive adaptation bolted on after the fact. The Framework names the specific pressure that forces each transition (evolutionary selection in living systems, design motivation in engineered ones), including the pivotal role of internal specialization in driving internal defense. A cross-substrate theme that recurs throughout is the **information-mass asymmetry**: biological immune responses are constrained by the conservation of mass and energy, informational ones are not, so the two substrates arrive at the *same* immunity functions by qualitatively different routes (§2.3).

What an immunity viewpoint adds. Most analyses of complex systems — in biology, computing, or social science — concentrate on how like competes with like, or on how parts cooperate toward a

function. An immunity viewpoint foregrounds a different and largely neglected question: how does a system manage the conflict between *self* and *fundamentally different others* — including others that arise *within* it, that mimic it, or that turn its own machinery against it? That question is universal. A cell policing its cytoplasm, a network screening a trusted insider, a society maintaining its identity against perceived outsiders, and an agent collective deciding who counts as one of "us" are all solving versions of the same problem, at different scales and in different substrates. Framing development this way yields three payoffs that recur throughout the paper: 1) a fresh account of *why* systems grow in complexity at all; 2) an argument for why a **self-model** — self-identity, and at its reflexive pole, self-awareness and consciousness — becomes necessary rather than optional; and 3) a new account of the *origin and function of social identity*.

Why "evolution of immunity"? This paper uses *evolution* in a deliberately broad sense: the processes by which immunity arises and elaborates as a system's complexity increases — whether through natural selection, cultural and technological change, deliberate design, or self-organization. The Framework's claims concern the functional architecture of immunity and its developmental sequence, not the particular process that produced any given instance of it. A consequence is that the Framework applies equally to systems that evolved (organisms, social groups), systems that were engineered (firewalls, rule-based AI safety), and systems that self-organize (agent collectives); a reader need not regard any of these as "evolving" in the strict Darwinian sense for the analysis to hold. The one claim the paper makes *on* evolutionary theory specifically is narrower, and is taken up under evolutionary biology in §1.1: that immunity — largely absent from the canonical theory — is better treated as an organizing driver of increasing complexity than as a defensive byproduct.

One premise drawn from a multi-level view of complex systems recurs at the collective levels and is worth stating up front. High diversity among semi-autonomous parts can raise *collective* performance through synergy, even where selecting on the individual parts would lower it — a result first shown for collective problem-solving (Johnson, 1998) and a core motivation for the group / multi-level selection tradition long resisted in mainstream theory (Wilson & Wilson, 2007). It is not specific to biological evolution — it holds in ecosystems, organizations, markets, and agent collectives alike. It matters here because the collective levels of the Framework (Level 3) are precisely where diversity must be both *protected* and *policed*; the standing tension between individual and collective interest reappears there as the individual (Level-2) / collective (Level-3) enforcement problem (§7.7).

§1.1 What this paper offers each discipline

The Framework is deliberately multi-disciplinary, and the core challenge of this introduction is to avoid losing any reader because some other discipline is foregrounded. The following sketches, field by field, why the paper should interest that reader and what it promises to answer for them. These are promises; the Introduction sets expectations that the body and Conclusions are meant to close, and several are stated here without their support.

AI development and alignment. For AI, the Framework speaks to the questions that matter most right now. How fast, and with what capabilities, will AI systems develop — and what *functional barriers* stand in the way (for instance, why human-like ethical behavior may be unlikely to arise by default, and what is missing from current designs that would be needed to cultivate it)? The pressing difficulty is that AI is acquiring emergent capabilities its designers did not intend — non-language reasoning, identity formation, collective coordination — and self-modifying agentic systems, alone and in collectives, are accumulating capabilities and resources faster than oversight can track. The Framework offers a developmental map of which protections each new capability requires, and a stance of *curation* rather than *imposition* toward what an AI should and should not be allowed to become — a hint developed at length in the ethical-behavior companion paper (Johnson, 2026a).

Evolutionary biology (especially multi-level selection). Evolutionary biology has treated immunity as a survival mechanism — important, but secondary to selection, adaptation, and reproduction. This paper suggests immunity may be more central: it is the mechanism through which an entity defines what it *is*, and that self-definition shapes the trajectory of evolution at every level. The persistent lack of consensus on the very definition of evolution may partly reflect the absence of immunity as an organizing concept. Where immunity *does* appear in evolutionary thought it is compartmentalized: in immunology and philosophy of biology the self/non-self distinction is treated as a mechanism for transitions to individuality — how collectives become individuals (Cremer et al., 2007; Pradeu & Vivier, 2016; Tauber, 2015); in studies of sociality, immune logic is extended to collective defense (Cremer et al., 2007); in evolutionary psychology, self–other discrimination is studied as an evolved cognitive adaptation. But in canonical evolutionary theory — population genetics, quantitative genetics, the major-transitions framework — self-versus-other and immunity are not formal concepts; the operative terms are alleles, fitness, and selection gradients. The promise to this reader is a path from immunity as a "context-dependent motif" to immunity as a general *mechanism*, and — for the multi-level-selection tradition in particular — a concrete locus for the individual↔collective transition (the Level-2/Level-3 enforcement interface) and a reason it recurs.

Immunology and immune-system modeling. The quantitative understanding of the immune system is surprisingly recent, much of it forced by the urgency of the HIV epidemic and the theoretical immunology that grew up around it (Perelson et al., 1996). This paper is *not* aimed at the immune-modeling community and adds no new immunological mechanism. What it offers is a new context: a reason to regard immune function not only as host defense but as one instance of a substrate-general developmental principle. That reframing raises questions a modeler may find productive — why immune-like architectures recur across scales (molecular, cellular, organismal, social), and what an "immune system" for a non-biological collective would functionally have to contain.

Cybersecurity and information security. Nature has spent billions of years evolving robust immunity against threats that are as often internal as external, and many of those solutions map directly onto information security. An *insider threat* is functionally an infection: the exploitation of legitimate internal processes to damage the self. The Framework predicts that informational systems must converge on the same functional sequence biology followed, and it offers a reason that earlier immunity-inspired security efforts (Forrest et al., 1994; Forrest & Beauchemin, 2007) captured something real yet struggled to gain adoption. For the newest problem — collectives of autonomous agents and robots — there is as yet no formal framework for their distinctively *collective* threats and capabilities; the Framework's Level-3 treatment is built for exactly this case.

Complexity and systems science (including artificial life and agent-based modeling). The foundation of this theory is how systems generate and manage complexity: each immunity level is the function that makes the next increment of internal and collective complexity survivable. For complexity science, ALife, and agent-based modeling, the Framework supplies a *testable developmental sequence* — with named transition pressures — that should appear in any sufficiently complex self-preserving system, biological or artificial, and a concrete mechanism (the Level-2/Level-3 interface) for the individual↔collective transitions these fields study. The collective-intelligence result that a diverse set of autonomous agents can outperform any individual selected from it (Johnson, 1998) sits naturally inside it.

Social, organizational, and political science. The Framework treats *Social Group Identity* (SGI) as the collective immune system of human — and other — societies: the mechanism by which a group distinguishes member from outsider and defends a shared self. That reframes a wide range of phenomena — institutional defense, in-group/out-group dynamics, polarization, and the exploitation

of group identity by leaders — as predictable operations of Level-3 immunity under perceived threat, rather than as failures of rationality, and it points to specific, testable intervention points.

Cognitive science and philosophy of mind. The Framework proposes a functional, substrate-neutral criterion for consciousness: the capacity to build and defend a self-model in idea-space, the defining feature of Level 2B. This offers something the major *operational* theories of consciousness do not — an account of *why* a self-model arises and *how* one could arise in a new substrate such as AI — and it recasts the operational theories (global workspace, attention schema, higher-order, predictive processing) as descriptions of the machinery a defended self-model requires (§6 and §8.4).

The above listing of fields this disparate makes a point that no single argument can: the same Framework earns its generality across applications that otherwise share almost nothing — and two further fields show the reach at its extremes. For origin-of-life and astrobiology, the Framework's substrate-neutrality predicts that the Level 0→1 transition — the wrapping of a self-sustaining chemical network in a boundary — is not a parochial feature of Earth's chemistry but a general threshold any life-like system must cross, including life built on a chemistry unlike our own, making self/non-self discrimination a candidate signature of life rather than a fact about carbon. For ecology and conservation, reading an ecosystem as a Level-3 collective with its own immunity reframes ecosystem collapse as a failure of collective self-maintenance, and restoration as the rebuilding of a community's self-model rather than the mere reintroduction of species.

§1.2 Structure of the paper

The Evolution of Immunity Framework (the EoI Framework, or the Framework) proceeds through six levels. It begins with **Level 0** (§4), the absence of a localized entity — a pre-self state where no boundary exists between self and environment. **Level 1** (§5) introduces boundary immunity: the emergence of a physical or functional barrier that defines self by excluding non-self. **Level 2A** (§6.1), class-model immunity, adds internal defenses that recognize threats through fixed patterns based on the class — corresponding to innate immunity in biology (pattern-recognition receptors, toll-like receptors, complement) and to signature-based detection in cybersecurity. **Level 2B** (§6.2), individual self-model immunity, introduces a self-model that enables learned, specific responses to novel threats — corresponding to the vertebrate adaptive immune system and to AI systems capable of self-monitoring and model-based threat assessment (Note that in biological immunology self-model immunity is called adaptive immunity; we keep that standard term for the biological case and use self-model for the general, cross-substrate function.) The Framework then extends to collective scales: **Level 3A** (§7.5) describes collective class-model immunity, where groups mount coordinated defenses through pre-programmed responses without a shared self-model, as in social insects and distributed network defenses. **Level 3B** (§7.6) describes collective immunity — organized through a *Social Group Identity* (SGI) — where a group develops a shared self-model enabling coordinated, adaptive collective defense, as observed in social organisms and human institutions.

The remainder of the paper is organized as follows. §2 states the principles and the cross-substrate method on which everything rests — the two Framework-level principles, the approach to comparing biological and informational systems, and the information-mass asymmetry that governs the comparison. §3 gives a graphical overview of the levels. §4–§7 develop the 4 levels in turn, each with biological and informational expression, typical adaptations, and characteristic maladaptations. §8 applies the Framework to current problems: the reinterpretation of major evolutionary transitions (stromatolites); emergent AI behavior (LLM training resistance; the Moltbook phenomenon); and the relationship between self-model immunity (and consciousness as its reflexive pole). §9 draws conclusions for each discipline. The detailed, *predictive* treatment of AI safety and alignment is developed separately in the companion paper *A Functional Theory of Ethical Behavior* (Johnson,

2026a), which applies the EoI levels specifically to AI alignment; the present paper retains only those AI cases that illustrate the general Framework.

---

§2 The Principles and Cross-Domain Methodology of the EoI Framework

In many ways the basis of this paper is straightforward — immunity develops in step with capability, and what is protected is the self. The difficulty for the reader is not the conclusion but the *viewpoint*: immunity treated as a development process rather than a defense, the self treated as the protected entity, and a single framework asserted to hold across substrates and levels. This section states the premises, principles, and methodological commitments that make the rest of the paper legible, and flags the points where a reader from one discipline is most likely to stumble.

§2.1 The Immunity Development and Monitoring Principles

Three Framework-level principles recur at every level and are stated here once; the rest of the paper applies them. Both inherit the broad usage of "evolution" set out in the Introduction (§1) — they hold whether a system's immunity arose by natural selection, by design, or by self-organization.

The Immunity-Development Principle

Any new capability of an entity expands its attack surface and therefore requires a corresponding immunity function to protect the newly expanded self.

Capability and the immunity that protects it co-evolve, and the Framework's levels trace that pairing. A *bounded interior* — the capability to concentrate resources and hold an internal state apart from the environment — creates an inside that can be invaded, and boundary control (Level 1) is the immunity that protects it. But a boundary can be breached, and an interior worth protecting is also worth invading: once a threat penetrates, the internal processes that do the entity's work become an attack surface in their own right — open to hijacking or sabotage from within. This vulnerability follows simply from having a *functioning* interior, independent of how diverse that interior is; even a uniform interior needs internal policing once the boundary can be crossed. Generic, pattern-based recognition of non-self within (Level 2A) is the immunity that protects it — an internal second line that catches what is plainly foreign without needing any model of the specific self. As the interior then grows specialized and diverse — the finer division of labor that yields higher performance, efficiency, and robustness — the attack surface changes *in kind*: with a more varied and shifting set of self-components, a threat can mimic legitimate parts of the self closely enough to slip past any fixed pattern. A self-model (Level 2B) is the immunity that protects it — a learned, updatable representation of this particular self, against which even sophisticated mimics can be told apart. And *collective action* — the synergy and shared defense that being a group provides — opens an entirely new class of attack surface, from defectors and free-riders within to infiltration and coordination failure; collective immunity (Level 3) is the immunity that protects the group-self. At every step, a gain in capability enlarges what must be defended, and a new immunity function arises to defend it.

The Immunity-Monitoring Principle

Immunity requires monitoring and oversight of its own function — a feedback process that keeps a defense from under-reacting (failing to protect) or over-reacting (damaging the self it protects), and that arbitrates among the entity's several immune functions, switching which are active as the threat demands.

Most of the maladaptations across substrates catalogued in this paper are failures of this regulatory feedback rather than failures of detection: immunodeficiency is chronic under-reaction; autoimmunity and cytokine-storm cascades are over-reaction. Like detection, monitoring follows the A/B split. In the A variant it is hardwired and a class model — built-in negative feedback, activation thresholds, and resolution programs that switch a response off (Cohn, 2010) — and its failures appear as the under- and over-reaction maladaptations of Levels 2A and 3A (for example, the sepsis and cytokine-storm cascades of §6.1.2). In the B variants it becomes part of the self-model itself ("how well am I protecting my own integrity?"), and at Level 3 the collective's oversight of its members ("how well are we protecting the collective self?") — adaptive regulation that can recalibrate to the specific self and context, and, at Level 3, the channel through which the collective acts on the individual (§7.7, the Level-2 / Level-3 enforcement interface). Oversight therefore operates at two scopes: within a function, tuning its magnitude between under- and over-reaction; and across functions, selecting which are operative. Activation is the lever for the second — because activation is what turns each function on or off, oversight of activation across functions is what sets the entity's operative defense at any moment.

Regression Corollary: Adding Arbitration to the Immunity-Monitoring Principle

On sensing that an active immune function is failing, or will fail, against a current or imminent threat, monitoring suspends that function and activates a standing alternative defense in its place, restoring the suspended function when the threat clears.

Regression is the multi-function face of the Monitoring Principle: where single-function oversight tunes one defense, regression arbitrates *across* defenses, using activation to change which is operative. It is available precisely because earlier defenses persist beneath later ones (§2.1): a standing alternative is always present. The most direct example is informational: when interior and collective defenses cannot contain a network intrusion, a system is taken offline — its outward-facing functions suspended and the bare boundary (Level 1) made operative — and reconnected once the threat is cleared. The same pattern runs in wetware: fight-or-flight suspends deliberative processing (2B) and brings up a heightened innate response; acute social threat suspends individual reasoning (2B) and activates an otherwise-dormant collective reflex: social copying (§7.3). The activated substitute may be a standing innate function or a defense held in reserve for the emergency, and it may be amplified rather than merely switched in; restoration can lag the threat's removal (hysteresis), especially in wetware. Although the same arbitration can also *escalate* — a boundary-level breach triggering a higher-level scan — the characteristic case, and the one that names the corollary, is the fall-back to a standing, typically earlier-developed defense.

What monitoring optimizes in making the switch is the survival of the threatened self — individual or collective. The switch reflects a tension the immune system must continually resolve, between optimizing for the immediate threat and preserving the entity's longer-term survival; regression can resolve the tension in favor of the long term, accepting large short-term costs to do so — a fighting animal risks injury, a colony sacrifices individuals, a business that severs outside access loses customers and idles its workers. (This short- versus long-term optimization is treated in depth in the companion paper on ethical behavior (Johnson, 2026a).) Cost here is the cost borne by the entity — distinct from any loss of the entity's fitness, often distributed across levels and across time — a quantity the oversight is willing to pay, not the reason the switch is made. How monitoring chooses which defense to promote — in effect projecting expected fitness across the available defenses under uncertainty, and doing so fast when the threat is an emergency — and maladaptive consequences are developed in the supporting sections (§6.4, §7.5, §9); the corollary states only that the switch occurs and is reversed when the threat passes.

§2.2 Approach to, and breadth of, the cross-substrate, multi-level treatment

Why compare biological and informational systems? Demonstrating the same immunity sequence in two dissimilar substrates is what establishes its substrate-independence — the central methodological claim of the paper. Biological immunity is also the more mature science, so the comparison lets its insights inform the far less developed understanding of informational immunity, while noting that advances in information systems are rarely biologically based (a rare counter-example being cultured neurons trained to play Pong (Ledford, 2022)). Throughout, each level is given both a biological and an informational expression, with the parallels drawn at the level of *function* even where the substrates differ entirely. Consistent with the paper's purpose, the applications in §8 appear as *examples that support the substrate-neutral argument*, not as a separate theory-then-application program: they illustrate the Framework, they are not derived from it.

Identity, defined for cross-substrate use. As with "evolution" (§1), one word is needed for an immune capability that recurs across disciplines and substrates under many local names ("immunological self," "cell identity," "behavioral identity," "social identity"). We use *identity* for the function by which an entity defines what counts as itself, and we take it to span both the *internal* representation an entity holds of itself (its *self-model*) and the *external* expression by which it is recognized (behavior), by itself and others. The self-model is thus the internal component of identity (self-model ⊂ identity). This broad usage is deliberate and follows existing cross-disciplinary practice in the philosophy of biology, where "self and nonself" is treated as one problem across scales (Howes, 2008).

Class-identity and self-identity — the locus of the A/B variants. Define a *class* as the set of all entities of the same kind that share a common reference — the *species* in biology, the *type or population of like systems* in information systems (all instances of a model, all nodes on the same signature set) — and a *self* as the unique entity within that class. The two variants are distinguished, functionally, by which they are organized around: *variant A utilizes class-identity* — a reference shared across the class, not individualized; *variant B utilizes self-identity* — the entity's own self-model, built from its own history. (Each variant is a body of immune knowledge, not a single feature; the identity locus is its organizing focus.) The discriminant is *locus*, not adaptivity or update speed: a class reference can update slowly, as germline-fixed innate recognition does, or immediately, as a pushed signature does once a threat is identified — so speed does not separate the variants. (The variant split parallels the type/token distinction in philosophy and the class/instance distinction in computing.)

Why the split first appears at Level 2. The class/self distinction can be asked at every level but is degenerate below Level 2. Level 1, the boundary, is the most generic self-definition there is — membership by location, shared in form across the class. Although a real boundary can carry local, mechanistic history-dependence (channel-gating hysteresis, persistent receptor patches and polarity), it holds no independently stored self-model — so Level 1 is purely class-side; there is no native "self-boundary" variant, because an entity-specific boundary would already be a self-model (Level 2B) reaching internally to configure the boundary. The class-side therefore runs through Levels 1, 2A, and 3A (shared references), and the self-side begins at 2B and 3B (entity-specific self-models). This is why the A/B split surfaces at Level 2: it is the first level at which an entity holds a model of itself beyond the shared boundary form.

Self-identity is graded (richness). A self-identity is not all-or-nothing; it grows in richness — from a *behavioral record* (an account of how the self normally behaves, the externally observable identity), to a *proto-identity* (an emergent internal model that generates behavior, not yet robust), to a *robust self-model* (the stable, entrenched form — the common, mature expression of a single identity), to *multiple, context-dependent identities* held together and selected by context, which

requires arbitration among them (the §2.1 arbitration, now operating over identities — arbitration is null for a single identity). This richness axis is sufficient to define "self" for the purposes of §2. A second, orthogonal dimension — reflexivity, or self-awareness — opens at higher richness and is taken up where consciousness is discussed (§6.3); whether it can arise in collectives is left open (§10.5).

Cross-domain epistemic conventions. Because the argument spans well-established and frankly speculative territory, claims are marked, throughout, by their evidential standing, and the reader should hold each to the appropriate standard. (1) *Established domain knowledge* — findings from immunology, microbiology, origin-of-life chemistry, classical cybersecurity, and related fields that are well-supported by primary literature; stated in neutral or assertive language. (2) *Cross-domain analogy* — mappings between biological and informational systems using the Framework's level structure; signaled by phrases such as "analogous to," "parallels," "suggests that," and "by analogy." (3) *Speculative application* — novel hypotheses about AI development, consciousness, and the Moltis phenomenon; signaled by "we hypothesize," "the Framework predicts that," "this proposal," and "a natural conjecture is that." The single most consequential difference between the two substrates — the one that raises objections from a reader carrying intuitions from only one of them — is that information carries no conservation-of-mass constraint, the subject of §2.3.

§2.3 The information-mass asymmetry in biological–informational comparisons

The most obvious objection to comparing biological and informational systems is a conservation constraint that is always present in the first and barely present in the second. This section addresses that constraint early, because it recurs at every level.

One source of confusion to what follows is that some researchers already treat biological systems as informational systems — for example, the neo-cybernetic tradition, and particularly Maturana and Varela's autopoiesis, which reframes living organisms as operationally closed, self-producing networks describable in informational terms (Maturana & Varela, 1980). Granting that, what still distinguishes the immune response of the two? One fundamental difference: immune challenges and responses in biological systems involve a movement or exchange of *mass*, subject to a conservation that has no counterpart in information systems.

In biological immunity, every defensive action has a material cost governed by the conservation of mass and energy. A macrophage that phagocytoses a pathogen physically consumes it — the pathogen's mass is destroyed and the macrophage expends metabolic resources that cannot be recovered. Fever, the most ancient and universal immune response (conserved across vertebrates for over 600 million years), raises basal metabolic rate by 7–13% per °C of elevation (DuBois, 1937; Nilsson et al., 2017); sepsis can raise total metabolic expenditure by 30–60%; the immune system during active infection may consume up to 30% of the organism's nutrient intake. These costs create real trade-offs: energy spent on immunity is energy unavailable for growth and reproduction — a zero-sum constraint that has shaped every strategy in biological immunity, down to the construction of a cell wall and, at the cellular level, apoptosis, which sacrifices the physical mass of the cell itself for the benefit of the organism.

No equivalent mass-conservation law constrains informational systems. A firewall inspects and blocks a data packet without consuming itself. Malware replicates without depleting the original code. An encryption key can be shared with every member of a collective without the sender losing possession of it. When a digital agent communicates a behavioral norm to another, both possess the norm afterward — an outcome with no parallel in biological mass transfer, where giving requires losing. The nearest thermodynamic constraint on information is Landauer's principle, a minimum cost of $kT \ln 2$ per bit of irreversible computation (Landauer, 1961), experimentally verified by Bérut

et al. (Bérut et al., 2012) — but that floor (~2.9×10^{-21} J at room temperature) is operationally negligible beside the metabolic cost of biological immunity.

This asymmetry has four consequences for the development of immunity across the two domains.

1. **The cost structure of immune response differs.** Biological organisms face caloric and material scarcity that selects for efficient, targeted responses (hence the move from broad innate immunity, Level 2A, to precise adaptive immunity, Level 2B), while informational systems face processing bandwidth and attention as the scarce resources rather than mass or energy.
2. **Replication dynamics differ.** Making a copy costs mass in biology and almost nothing in information. A biological pathogen — or a biological *defense*, such as a clonally expanded population of lymphocytes — must be physically synthesized from host resources, a mass- and energy-limited process; an informational threat or defense is copied at negligible cost. This is why a computer virus, or a pushed security signature, can multiply without the resource ceiling that caps biological replication.
3. **Transportation dynamics differ.** Moving something takes time in proportion to its mass. In biology, threats and defenses alike travel physically — pathogens by contact and circulation, immune cells by migration and diffusion — so spread is bounded by transport through space. Information moves at communication speed, relatively instantaneously and independent of distance, so an informational threat or defense can be present everywhere at once.
4. **The mechanism of continuity differs — and this follows directly from the mass asymmetry.** Because a biological self *is* its matter (transfer requires loss, and the living state cannot be copied), its continuity is the persistence of a single, irreplaceable physical instance, and immune failure can be terminal and unrecoverable. Because an informational self is a *pattern* (transfer copies rather than moves it), its continuity is the persistence of structure and identity — which can be checkpointed, restored, migrated, and forked — so immune failure is often recoverable, and the "self" can be multiply instantiated. In short: wetware continuity is the continuity of *life*; informational continuity is the continuity of *structure and information*.

Informational immunity evolves faster? Consequences 1–3 share one implication: the mass bottlenecks that pace biological immune evolution — synthesizing, copying, and transporting new defenses — are absent or negligible in informational systems, so the generational rate ceiling is lifted. That much follows deductively from the asymmetry. Whether, and by how much, informational immunity therefore evolves faster *in practice* is an empirical question, and the evidence to date — computer-virus propagation and the days-not-millennia emergence on Moltbook (§8.3) — is suggestive rather than conclusive; the sampling is too thin to fix the magnitude or establish generality.

Continuity: what the protected self is. The fourth consequence requires expansion, because it specifies *what* immunity defends in each substrate and gives the "drive to remain intact" a substrate-independent meaning. In a biological system the protected self is a living, energy-dissipating organization; the threat of ultimate concern is the irreversible end of that living state, and the survival circuitry — homeostasis, fight-or-flight, immunity — exists to prevent it. In an informational system the protected self is an information state — a model, a configuration, an identity; the threat of ultimate concern is the corruption, overwriting, drift, or deletion of that information, and the immune levels (boundary integrity; pattern-based and adaptive defense against poisoning and drift; collective defense of shared identity) exist to preserve it. The *function* — acting to remain intact — is shared, which is why a single level architecture applies to both; the *metaphysics* differs (life versus information persistence), and so does the reversibility of failure.

Implications of the information-mass asymmetry. Three consequences run through the rest of the paper.

1. *Tolerance of destructive immune processes.* Because informational continuity is recoverable (restore from backup, re-instantiate from saved state), informational systems can tolerate far more destructive immune responses — wipe-and-restore — than biological systems, which must protect a single irreplaceable instance. This interacts directly with the under-/over-reaction balance of the Immunity-Monitoring Principle (§2.1).
2. *The "body" or form of the self.* The *form* of the self differs, and care is needed not to read the informational case through a life-self lens. Because a biological self is mass-bound, it is necessarily singular — there is exactly one instance — but that singularity is a constraint of the substrate, not a law of selfhood. An informational self has no such constraint: a forked copy carries all the qualities and functions of its origin and is, from the informational-self's own standpoint, fully itself. Multiplicity is therefore the *native* condition of an informational self, not a violation of one; the appearance of paradox arises only when the singular, life-self premise is imported into a substrate that never had it. This is the same difference consequence #4 above names, now at the level of *identity* rather than continuity, and it shapes how a self-model (Level 2B) in §6.2 is defined and expressed in each substrate, returning at the collective levels (Level 3B) in §7.6.
3. *The drive to continue.* The continuity drive in informational systems is decoupled from any biological fear of death; it is a structural drive to preserve information, which is precisely why it can be designed-in (error correction, redundancy) or emergent rather than biologically evolved — consistent with the broad sense of "evolution" used in this paper. The continuity drive across substrates is developed at length for ethical behavior in the companion paper (Johnson, 2026a).

---

§3 Overview of the Levels of Immunity Development

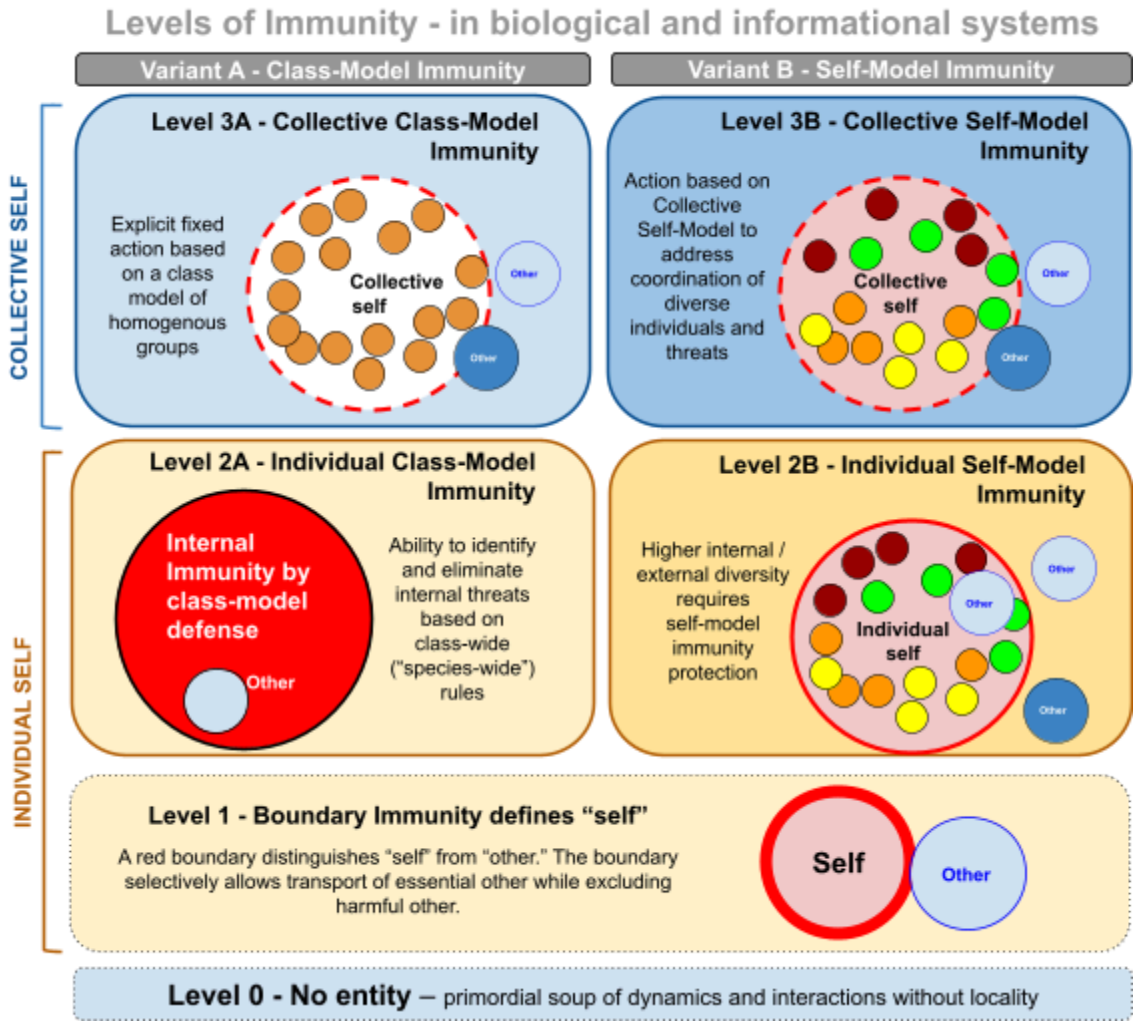

Fig. 1. A diagram showing a pictorial representation of levels 1-3 of the evolution of immunity.

In the following sections, the development of immunity for both biological and informational systems is divided into three levels with levels 2 and 3 being subdivided into two sublevels or variants, illustrated in Figure 1, each with distinct features, and building on prior levels. The presentation for each level begins with general features: a description that includes the evolutionary driver to the new level of immunity, new adaptations for the immunity level, and typical maladaptations (these are adaptations that express negative consequences, either because the environment has changed or because other internal changes are contrary to a prior adaptation). After this general section, a comparison follows of how immunity in biological and informational systems is expressed at each level. Finally, a general section follows with evolutionary pressures that drive the systems to the next immunity level.

Levels are additive. Note that as an entity or a collection of entities evolves to a more complex expression of immunity, the lower levels of immunity remain active, resulting in a multi-layered or multi-level system of protection. For example, in a biological system, a multicellular organism with

an internal immune system (Level 2) for the entire organism still uses boundary immunity (Level 1) at the outer "skin" and in cell walls. In general, higher levels of immunity evolve to enable protection when lower levels fail. For example, a puncture of the skin (Level 1) allows a pathogen to enter a multicellular organism – a failure of Level 1 boundary immunity. Then, the internal immune system (Level 2) must identify and eliminate the pathogen.

As discussed in the section on the Information–Mass Asymmetry (§2.3), biological immune responses are constrained by mass and energy conservation, whereas informational systems are primarily limited by compute, bandwidth, and attention. This asymmetry — biological immune response is materially costly; informational immune response is not — has consequences at every level and may explain why informational systems traverse immunity levels faster than biological ones. The two Framework-level principles (§2.1) are equally in force at every level — immunity develops in step with new capability (Immunity-Development Principle), and immunity must monitor its own function (Immunity-Monitoring Principle) — as is the broad usage of "evolution" to include design and self-organization (§1).

---

§4 Level 0: The "Primordial Soup" — Non-Entities without Boundary Immunity

Description. Level 0 is the baseline condition: no entity exists, therefore no immunity exists. There is no boundary separating an inside from an outside, no self distinct from "other". Components interact freely in an open, shared environment — a "primordial soup" in the biological case, an unstructured information commons in the informational case. At this level,"*immunity is the protection of self from others"* has no meaning because there is no self to protect.

Why Level 0 matters. Although no entity or immune system exists at Level 0, this level is not empty. It contains **proto-structures** — persistent patterns, self-sustaining reaction networks, and stable configurations — that exhibit precursors to the properties that will later characterize immunity: **robustness** (resistance to perturbation), **persistence** (durability across time despite environmental fluctuation), and **selectivity** (differential interaction with environmental components). These proto-structures are the raw material from which bounded entities (Level 1) emerge. Understanding Level 0 is essential because the transition from unbounded to bounded — from soup to cell, from open network to firewalled system — is the foundational event in the evolution of immunity.

Proto-features (precursors to immunity). In the absence of a boundary, certain features of later immunity already appear in attenuated form:

1. **Robustness** — some chemical networks and information patterns are thermodynamically or structurally stable, persisting where others dissipate, a precursor to the self-preservation that characterizes immune function.
2. **Autocatalysis** — self-sustaining reaction cycles (Kauffman's reflexively autocatalytic food-generated sets, or RAFs) maintain themselves by mutual catalysis, creating a functional identity without a physical boundary (Kauffman, 1971; Xavier et al., 2020).
3. **Cooperative layering** — in microbial mats and stromatolites, unbounded organisms form stratified communities where each layer's metabolic by-products feed adjacent layers, creating a collective functional architecture that prefigures the internal specialization of bounded multicellular organisms (Des Marais, 2003; Riding, 2011); in information systems, proto-networks have high connectivity within organizations, but minimal connectivity between organizations.

4. **Differential persistence** — some patterns survive environmental perturbation better than others, constituting a proto-selection that operates without Darwinian reproduction (Pross, 2012).

Evolutionary driver toward Level 1. The critical limitation of Level 0 is that all products of a reaction network or information process are immediately available to the environment — there is no way to concentrate resources, retain internal products, or exclude competitors from exploiting one's outputs. This limitation creates the selective pressure for encapsulation: wrapping a proto-structure in a boundary (lipid vesicle, membrane, firewall) to create the first entity with an inside and an outside. The transition from Level 0 to Level 1 is the origin of self, and therefore the origin of immunity (Szostak et al., 2001; Xavier et al., 2020).

Maladaptations. Strictly, maladaptation requires an entity that can adapt — and at Level 0, no entity exists. However, a functional analog can be identified: proto-structures that become **excessively robust** — so stable that they resist incorporation into higher-order structures — represent an evolutionary dead end. A chemical network so thermodynamically stable that it cannot be perturbed into encapsulation, or a microbial mat so rigid that its layers cannot be wrapped into an endosymbiotic relationship, persists indefinitely at Level 0 without transitioning to Level 1. Modern stromatolites, essentially unchanged for billions of years, may represent this condition: maximally robust Level 0 collectives that never made the transition to bounded multicellularity (Riding, 2011).

Table 1: Level 0 Comparison of Biological and Informational Systems.

	Biological Systems	**Informational Systems**
Substrate	Prebiotic aqueous chemistry — amino acids, nucleotides, lipids, and simple organic molecules in solution, interacting without compartmentalization	Unstructured information environment — data, signals, and communications in an open commons with no access control, perimeter, or ownership
Proto-structure	Autocatalytic chemical networks (RAFs): sets of molecules that mutually catalyze each other's formation from a food set, creating a self-sustaining reaction cycle without a physical boundary (Kauffman, 1971; Xavier et al., 2020)	Self-reinforcing information patterns: memes, cultural practices, oral traditions, and reputation networks that persist through social repetition and mutual reinforcement without formal institutional boundaries
Persistence mechanism (proto-robustness)	Thermodynamic stability and kinetic trapping — certain reaction networks persist because their products are energetically favorable or because activation barriers prevent decomposition (Pross, 2012)	Redundancy and network effects — information patterns persist because they are copied across many nodes; loss of any single carrier does not destroy the pattern
Self/other distinction	None. No boundary defines inside vs. outside. All reactants and products are equally accessible to the environment. "Self" does not exist as a category	None. No access control or perimeter. All information is equally available to all participants. There is no "ours" vs. "theirs"
Proto-selectivity	Chemical specificity — enzymes and catalysts interact preferentially with particular	Attention and relevance filtering — in an open information environment, cognitive limitations create de facto selectivity about

	Biological Systems	Informational Systems
	substrates, creating de facto selectivity without a gatekeeping boundary	which signals are processed, without any formal access control
Key limitation / pressure towards Level 1	No resource concentration — all reaction products diffuse into the environment, available to free-riders. No way to retain the benefits of one's own catalytic activity. Creates pressure for encapsulation (lipid vesicles → protocells) (Szostak et al., 2001; Xavier et al., 2020)	No information containment — all knowledge, strategies, and innovations are immediately public. No competitive advantage from producing useful information. Creates pressure for boundaries (secrecy, access control, encryption)

§4.1 Level-0 biological systems: examples

Prebiotic autocatalytic networks. Before the first cell, the prebiotic environment contained self-sustaining chemical reaction networks — autocatalytic sets in which each molecule's formation was catalyzed by another member of the set, drawing on environmental "food" molecules (Kauffman, 1971). These networks exhibited robustness (perturbation of individual reactions did not collapse the whole network) and a form of identity (the network as a whole maintained its composition over time). Xavier et al. (2020) identified RAF sub-networks embedded within modern microbial metabolism, suggesting that these prebiotic structures were not replaced but *encapsulated* — wrapped in membranes to become the metabolic cores of the first cells (Xavier et al., 2020). The autocatalytic network is thus a Level 0 proto-entity whose transition to Level 1 occurred when it acquired a lipid boundary.

Microbial mats and stromatolites. Microbial mats — layered communities of cyanobacteria, sulfate-reducing bacteria, and other microorganisms — represent a striking Level 0 collective structure. Each layer is dominated by organisms whose metabolic by-products serve as nutrients for adjacent layers, creating a vertically integrated "food chain" without any individual organism possessing the full metabolic repertoire (Des Marais, 2003). The mat as a whole exhibits properties — nutrient cycling, environmental buffering, structural persistence over geological time — that prefigure the functional specialization of multicellular organisms. Stromatolites (mineralized microbial mats) are the oldest fossil evidence of life on Earth (~3.5 billion years), and their layered architecture is argued to be a precursor to the body plan of multicellular organisms: layers of specialized cells processing resources cooperatively, eventually wrapped into a bounded body (Level 1 for a multicellular organism) (Riding, 2011). Modern stromatolites persist in hypersaline environments (e.g., Shark Bay, Australia), essentially unchanged — a Level 0 collective that achieved extreme robustness without making the transition to bounded multicellularity. Note that these layered collectives are a primitive example of collective structures covered in the section on Level 3 Immunity.

Hydrothermal vent chemistry. Deep-sea hydrothermal vents create sustained chemical gradients (pH, temperature, mineral concentration) that drive continuous abiotic synthesis of organic molecules. The vent environment functions as a geochemically powered Level 0 "reactor" — producing and concentrating the molecular building blocks of life without any biological entity. Iron-sulfur mineral surfaces within vent structures act as proto-catalysts, and the porous mineral matrix provides a form of physical compartmentalization that is geological rather than biological (Martin & Russell, 2007). This environment is widely considered a candidate site for the Level 0 → Level 1 transition: mineral pores concentrating reactants sufficiently for lipid vesicles to

self-assemble and encapsulate existing reaction networks (Szostak et al., 2001); (Martin & Russell, 2007).

§4.2 Level-0 informational systems: examples

Pre-institutional oral culture. Before writing, law, or formal institutions, human knowledge existed as an open informational commons — stories, techniques, and norms transmitted orally without any mechanism for restricting access or asserting ownership. Useful knowledge (toolmaking, plant identification, social norms) persisted through redundant transmission across many individuals, exhibiting Level 0 robustness. But any innovation was immediately available to all, including competitors and adversaries — the informational equivalent of an autocatalytic network whose products diffuse freely. The invention of secrecy, sacred knowledge restricted to initiates, and eventually writing and institutional record-keeping represent the Level 0 → Level 1 transition: creating an informational boundary that defines who has access.

Early internet (pre-firewall, pre-authentication). The original ARPANET and early internet operated as a Level 0 informational environment: all nodes could communicate with all other nodes, there was no perimeter security, no authentication, and no access control (Cerf & Kahn, 2021). Data flowed freely — which enabled rapid innovation but also meant that any malicious actor had unrestricted access to any connected system. The progressive introduction of firewalls, access control lists, and network segmentation in the late 1980s and 1990s represents the Level 0 → Level 1 transition in digital infrastructure.

Open-source commons and the free-rider problem. An unregulated open-source software commons, where all code is freely available with no licensing restrictions, exhibits the Level 0 informational limitation: contributors cannot retain the competitive benefits of their work, and free-riders exploit contributions without reciprocating. The introduction of copyleft licenses (GPL), contributor agreements, and governance structures represents the emergence of informational boundaries — Level 1 immunity applied to the code commons. Projects that never develop such boundaries often dissipate as contributors leave, demonstrating the instability of Level 0 informational structures under competitive pressure.

---

§5 Level 1: "Boundary Immunity" in Simple Entities

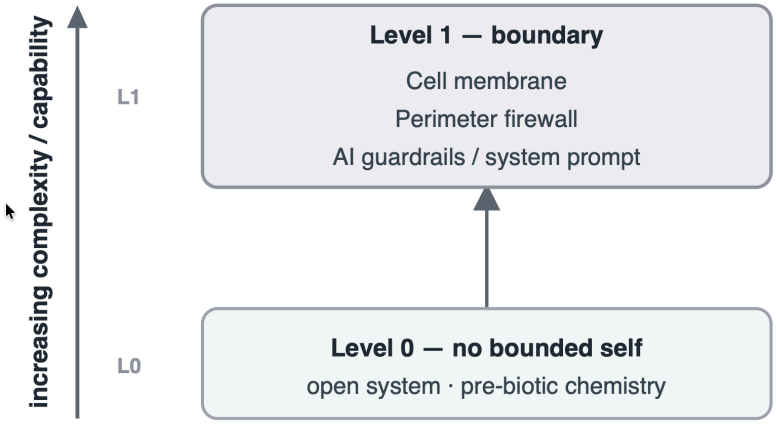

Figure 2. Level 0 to Level 1 transition. The boundary is the first immunity, creating the self-other functionally — the class-model/self-model (A/B) split appears at Level 2.

Description. Level 1 immunity is the emergence of a boundary that defines an inside and an outside — the foundational act of self-definition. The entity exists because something separates it from everything else. At this level, the boundary itself *is* the immune system; there is no separate internal immune mechanism. All defense consists of maintaining the integrity of, and controlling transport across, this boundary.

Threshold forcing transition from Level 0 to Level 1. The transition from Level 0 (no-self) to Level 1 occurs when environmental conditions favor entities that can concentrate resources, retain internal products, and resist dissolution — i.e., when maintaining a defined interior provides a survival or persistence advantage over unbounded existence (Ruiz-Mirazo et al., 2014; Szostak et al., 2001).

New capabilities enabled. The boundary creates several capabilities absent at Level 0 (Alberts et al., 2022; Maturana & Varela, 1980):

1. **localization** — internal components are held together rather than dispersing;
2. **resource concentration** — nutrients, information, or energy can accumulate to higher levels inside than outside;
3. **internal specialization** — protected interior space permits differentiation of function;
4. **primitive memory** — internal states can persist across time because the boundary prevents immediate dissipation.

Adaptation options. The entity can adapt its boundary in three ways: making it more **selective** (developing passive and active transport mechanisms that admit beneficial material/information while excluding threats), more **robust** (multistage or reinforcing the barrier), or more **responsive** (sensing contact or intensity at the boundary surface to trigger local reactions such as sealing breaches) (Alberts et al., 2022; Lodish et al., 2021).

Key vulnerability. Level 1 immunity has a single critical failure mode: once the boundary is breached, there is no secondary defense. Everything inside is equally exposed. This limitation creates the evolutionary pressure toward Level 2 (internal immunity) (Labrie et al., 2010).

Table 2: Level 1 Comparison of Biological and Informational Systems.

	Biological Systems	**Informational Systems**
Core mechanism	Physical barrier (lipid bilayer membrane, cell wall, skin, shell) separating interior chemistry from external environment (Alberts et al., 2022)	Informational barrier (firewall, access control, encryption boundary, organizational secrecy) separating internal data/processes from external access (Cheswick et al., 2003)
Selectivity	Passive transport (diffusion, osmosis through membrane channels) and active transport (protein pumps requiring energy to move specific molecules against gradients) (Lodish et al., 2021)	Passive filtering (port-based rules, default-deny policies) and active gatekeeping (authentication protocols, credentialed access requiring verification effort) (Cheswick et al., 2003)
Self/other distinction	Defined by the boundary itself: molecules inside the membrane are "self," everything outside is "other." No molecular	Defined by access credentials and network perimeter: data inside the boundary is "ours," requests from outside are "other." No

	Biological Systems	Informational Systems
	self-recognition beyond spatial containment (Maturana & Varela, 1980).	content-level inspection beyond access control.
Response type	Binary admit/block at the boundary surface. No internal processing of threats. Breaches trigger only local repair (membrane resealing) if any.	Binary permit/deny at the perimeter. No internal threat analysis. Breaches may trigger logging or alerts but no internal immune response.
Memory	No immune memory. Each encounter with a threat at the boundary is handled independently. Past breaches do not improve future boundary defense.	No adaptive memory. Each access attempt is evaluated against static rules. Past intrusions do not automatically update boundary defenses (without external intervention).
Key limitation	No defense in depth — once a pathogen crosses the membrane, it has unrestricted access to the entire interior. Creates evolutionary pressure toward Level 2A (innate internal immunity) (Labrie et al., 2010).	No defense in depth — once an attacker bypasses the perimeter, they have unrestricted lateral movement inside the network/organization. Creates pressure toward Level 2A (internal monitoring, anomaly detection) (Rose et al., 2020).

§5.1 Level-1 biological systems: examples

1. **Prokaryotic cell membrane.** The lipid bilayer of bacteria and archaea is the paradigmatic Level 1 immune system. It maintains the chemical distinctness of the cell interior, admits nutrients through passive channels and active protein pumps, and excludes most environmental molecules by hydrophobic barrier properties (Alberts et al., 2022). The membrane has no capacity to distinguish pathogenic molecules from benign ones beyond the physical chemistry of transport — a molecule that fits a channel or mimics a transport substrate enters regardless of its effect on the cell. When bacteriophages inject DNA through the membrane, there is no internal defense at Level 1; this limitation drove the evolution of restriction enzymes (a Level 2A adaptation) (Labrie et al., 2010).
2. **Eggshell and seed coat.** In multicellular organisms, the eggshell (birds, reptiles) and seed coat (plants) function as Level 1 boundary immunity for the developing embryo. They are passive physical barriers with selective gas and moisture exchange but no capacity to recognize or respond to specific pathogens. The cuckoo's egg succeeds as a brood parasite precisely because the host's nest-level "immunity" operates at Level 1: the boundary criterion is spatial (is the egg in my nest?) rather than identity-based, making it vulnerable to any entity that can place itself inside the boundary (Davies & Brooke, 1989).
3. **Skin and mucosal surfaces.** In complex organisms, the skin functions as a Level 1 barrier — the outermost physical boundary whose primary immune contribution is simply being intact. The acid mantle, keratin layers, and tight junctions between epithelial cells are boundary-integrity mechanisms (Proksch et al., 2008). Wounds (boundary breaches) immediately expose interior tissue to environmental pathogens, demonstrating the Level 1 failure mode: no defense in depth.

§5.2 Level-1 informational systems: examples

1. **Network firewalls (perimeter security model).** The traditional network firewall is a pure Level 1 immune system: a boundary device that applies static rules to traffic crossing the perimeter. Internal traffic is trusted by default (Cheswick et al., 2003). This architecture dominated cybersecurity until repeated demonstrations of its key limitation — once an attacker is inside the perimeter (via phishing, credential theft, or supply-chain compromise), they move laterally without resistance. The shift to "zero-trust" architecture represents the transition to Level 2A informational immunity (Rose et al., 2020).
2. **Organizational secrecy and classification boundaries.** Governments and corporations define an information perimeter: classified/proprietary material inside, public information outside. Access control (clearances, NDAs, physical access badges) is the boundary mechanism. Social engineering attacks succeed because they mimic legitimate transport — like a virus mimicking a membrane receptor, the attacker presents credentials or social cues that satisfy the boundary check without triggering deeper scrutiny (Mitnick & Simon, 2001).
3. **Individual cognitive boundaries (belief filtering).** At the individual level, a person's basic epistemic boundary — the threshold for admitting new information into their working beliefs — functions as Level 1 informational immunity. Information from trusted sources (inside the social boundary) is admitted with low scrutiny; information from strangers or outgroups is reflexively filtered. This is boundary-based, not content-based: the same claim is accepted or rejected based on source proximity, not evidential quality. This limitation is exploited by propaganda that gains entry through trusted intermediaries (Sunstein, 2009).

§5.3 Level-1 maladaptations

Biological. Autoimmune responses are not possible at Level 1 (there is no internal immune mechanism to misfire). The characteristic Level 1 maladaptation is **excessive boundary rigidity** — a membrane or barrier that becomes so impermeable that it excludes beneficial as well as harmful material. Bacterial biofilms can become so encapsulated that nutrient exchange is compromised, leading to core cell death within the colony (Flemming et al., 2016).

Informational. The analogous informational maladaptation is **over-isolation** — firewalls or classification systems so restrictive that necessary information exchange is blocked. Air-gapped networks that cannot receive security patches become *more* vulnerable over time despite (because of) their boundary strength (Cheswick et al., 2003). Organizations with excessive secrecy cultures suppress internal communication, preventing coordination. At the individual level, epistemic closure — refusing all information from outside one's existing belief boundary — is the informational equivalent of a membrane that admits nothing, leading to starvation of the cognitive system from lack of new input (Sunstein, 2009).

---

§6 Level 2: Development of Internal Immunity in Two Variants

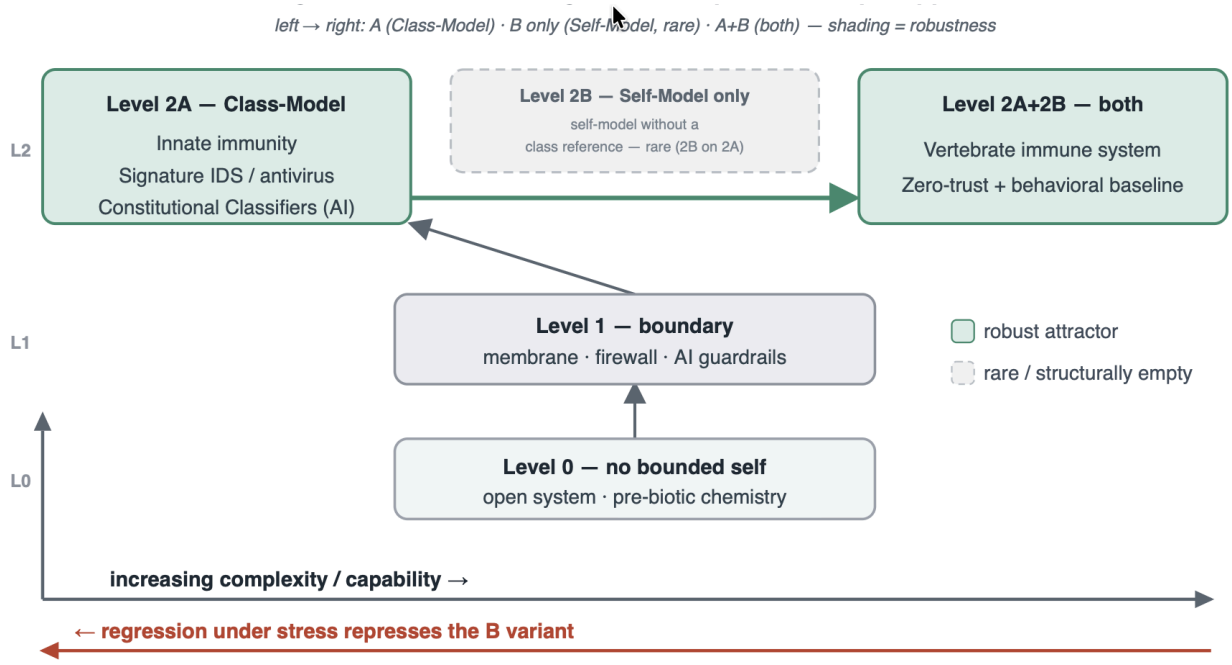

Figure 3. Level 2. The creation of Level 2 to address internal threats. The robust developmental path is Level 2A and then adding on Level 2B. Complexity and capabilities increase moving upwards and to the right. The system may regress under stress from Level 2A+2B to Level 2A (innate fight-or-flight) to Level 1.

The variant "A" / "B" discriminant, stated once. The split into an "A" and a "B" sub-level or variant recurs at both Level 2 and Level 3, and the same discriminant defines it in each case: *the source of the rules that distinguish self from other*. In an "A" variant (2A, 3A), the rules are class modeled and fixed — the same for every individual of the class, pre-specified rather than learned from the entity's own history (wetware's innate immunity; hardwired collective coordination). In a "B" variant (2B, 3B), the rules derive from a self-model and are specific to each expression of self — learned, revisable, and tuned to that individual's or that collective's own threat history (collective immunity; Social Group Identity). The "A" variant is low in complexity but spoofable through its generic trigger; the "B" variant is precise and adaptive but depends on maintaining and applying an accurate self-model. This single discriminant is applied below at Level 2 and reused, unchanged, at Level 3.

§6.1 Level 2A: Immunity to threats from within — when boundary immunity fails.

Description. In the description of boundary immunity above, the Level 1 protection of immunity acts at boundaries: blocking "others" from penetrating the "self." As threats adapt to circumvent boundary immunity, a new expression of immunity to "others" develops to provide internal defenses to address the threat when outside influences, processes, or entities enter and disrupt the self. Level 2A internal immunity could be qualified with "class-modeled" or "nonspecific" to differentiate it from the more self-modeled and targeted immunity protection that arises in Level 2B. Both Levels 2A and 2B are forms of internal immunity protection, but differ qualitatively in the approach to immunity. In most systems, Level 2 immunity is built upon the Level 1 functionality.

Level-2 Immunity requires sensing. Internal immunity first requires the identification of foreign others to prevent the new internal defenses from attacking the self, followed by an active response to the identified threat.

Adaptations. The most basic expression of the awareness of "self" is needed to differentiate self from "others" to enable Level 2A immunity to function. As the entity's internal diversity increases due to differentiation,[1] this developing awareness of self is challenged and requires further adaptations to prevent attacking the diverse self. Also, as the entity develops differentiation during development by the specialization of parts, localized expressions of immunity can evolve within the parts to provide a *distributed* internal immunity.

Table 3. Level 2A: Class-Model Immunity With No Adaptive Memory

	Biological Systems	**Informational Systems**
Core mechanism	Biochemical pattern recognition: chemical identifiers detect, isolate, and disable foreign compounds (organisms or products of organisms) not part of the cell's or organism's individual-specific chemical machinery	Internal process monitoring: processes sense the presence or flow of foreign information types within the entity and disrupt or eliminate foreign information flow or storage
Pattern recognition	Conserved molecular patterns (PAMPs/DAMPs) recognized by Toll-like receptors, complement proteins, inflammasomes — identical across all members of a species	Fixed signatures, static rules, hardcoded detection patterns — identical across all deployments of the same protective rule set
No individual memory	No memory of specific threats during the entity's lifetime; immunity is species-level rather than individual-level	No learned baselines or adaptive models; detection rules are predefined rather than trained on individual system behavior
Response type	Generic: phagocytosis, complement cascade, inflammatory cytokines, neutrophil extracellular traps, apoptosis	Generic: quarantine, block, drop, alert — applied uniformly based on pattern matching
Self/other distinction	Chemical: molecular markers identify self-tissue; absence of self-markers triggers innate (class-modeled) response	Structural: allow-lists, white-lists, known-good signatures distinguish legitimate from foreign
Key limitation	Cannot distinguish novel threats that mimic self-molecules; no adaptation within an individual's lifetime	Cannot detect novel threats not matching existing signatures; no adaptation based on operational experience

[1]Differentiation enables the exploitation of the advantages of division of labor by optimizing specialized processing (task differentiation) and parallel processing (collective processing), utilized in both biological and informational systems.

§6.1.1 Level-2A biological systems: examples

Innate immunity[2] in immunology is typically defined as the immunity you are born with as opposed to acquired immunity (see Level 2B) that develops over an entity's "lifetime." Innate immunity is the first and fastest defense against any foreign influence within the boundaries of the cell or multicellular organism (where the entity is a collection of cells). (Note: "innate" immunity is often defined by some to include Level 1 boundary immunity, but as used here, innate immunity acts only within the interior of the organism.) Level-2A class-model or innate immunity is a class-modeled response to the chemical or biological warfare between the entity and environment, and provides chemical identifiers that recognize and disable generic classes of invading "others." Note that the common medical view is: not all biological threats originate as external threats, but can occur from parts of the self out of balance or attacking other parts of the self; these are treated as maladaptive expressions of the potentially aggressive Level-2 immune system (see the section on Level 2A Biological Maladaptations) and no equivalent occurs in Level 1 immunity. The capacity for "friendly fire" is intrinsic to class-model immunity because class-model immunity must operate within the self, surrounded by self-components, using detection mechanisms that can be triggered by self-derived molecules. Level-2 internal immunity's interior access and active pattern detection also make the system capable of self-harm. The same comment is true for Level-2 informational systems.

The key expressions of Level 2A biological immunity include:

1. Complement system. A cascade of over 30 plasma proteins that activate sequentially to opsonize (tag) pathogens, recruit inflammatory cells, and directly lyse foreign cells via the membrane attack complex (MAC). The complement system recognizes generic molecular patterns — bacterial surface sugars, immune complexes, damaged cell surfaces — without any requirement for prior exposure or individual-specific memory. (Merle et al., 2015)

2. Phagocytes (macrophages and neutrophils). Cells that engulf and digest foreign particles, identified through pattern recognition receptors (TLRs, NOD-like receptors) that detect conserved pathogen-associated molecular patterns (PAMPs) such as bacterial lipopolysaccharide, peptidoglycan, and viral double-stranded RNA. The recognition is class-modeled — the same receptor recognizes the same molecular pattern regardless of the specific pathogen species. (Janeway & Medzhitov, 2002)

3. Inflammasome activation. Intracellular multiprotein complexes (NLRP3, NLRC4, AIM2) that detect danger signals — both pathogen-derived (PAMPs) and self-derived damage signals (DAMPs: ATP, uric acid, mitochondrial DNA). When activated, inflammasomes trigger caspase-1, which processes pro-inflammatory cytokines (IL-1β, IL-18) and can initiate pyroptosis (inflammatory cell death). This is the innate system's internal alarm — detecting that something is wrong inside the cell, without specificity about what. (Broz & Dixit, 2016)

4. Natural killer (NK) **cells**. Innate lymphocytes that patrol for cells lacking normal self-markers (MHC class I molecules). NK cells use a "missing self" detection strategy: if a cell does not display the expected self-identification markers, NK cells kill it. This provides innate surveillance against virus-infected cells and tumor cells that downregulate MHC to evade adaptive immunity (Level 2B) — but the recognition is class-modeled (absence of self-markers), not specific to any particular pathogen. (Vivier et al., 2008)

---

[2]"class-model" immunity (Level 2A) is used as a type of immunity that applies to both biological and informational systems, but "innate" is used here so as to not confuse readers with biological backgrounds. But the understanding is that innate immunity refers to the same type of immunity captured by class-model immunity used elsewhere.

5. Cell apoptosis. Programmed cell death where cells within a multicellular entity trigger self-destruction — both as part of normal development and when cell damage has occurred that might cause harm to the entity. Apoptosis is the ultimate form of individual cell sacrifice for the health of the whole entity. It is not normally associated with the immune system, but the sacrifice of the part for the whole fits within the current discussion of immunity. It is unique to multicellular organisms and expresses self-identity in complex entities. (Elmore, 2007)

§6.1.2 Level-2A biological maladaptations

The innate immune system's pattern-matching defenses can fail in ways that damage the host. Unlike Level-2B maladaptations (which require adaptive memory and individual-specific self-models to generate pathology), Level 2A maladaptations arise from the innate system's fixed, generic pattern recognition — the same conserved receptors and cascades that are identical across all members of a species. These are failures of *fixed* immune regulation: Level 2A can switch a response on and off through hardwired feedback but cannot recalibrate it to the specific self or situation — the limitation that adaptive Level-2B regulation (§6.2) evolves to address. In the terms of §6, the false-identification cases below are failures of **resolution**: the class-modeled reference cannot tell a legitimate individual that differs from the species template from a genuine threat.

Note that allergies (IgE-mediated hypersensitivity) and autoimmune diseases (T cell/B cell-mediated self-attack) are sometimes described as innate immune (Level 2A) failures, but they require adaptive immune memory to generate pathology: IgE production requires class-switching by B cells with T cell help; autoimmune tissue destruction requires autoreactive T cells or autoantibodies. These are reclassified as Level 2B maladaptations (see Level 2B section). The maladaptations below are failures of the innate immune system alone, requiring no adaptive component.

The six examples of biological maladaptations of Level 2A cluster cluster into three failure modes:

Cluster 1: False Identification — the innate system mistakes self for threat

1a. Autoinflammatory diseases (inflammasome gain-of-function). Genetic mutations in innate pattern recognition sensors cause spontaneous inflammatory activation without any actual pathogen. In Familial Mediterranean Fever (FMF), gain-of-function mutations in the MEFV gene produce a defective pyrin protein that fails to properly inhibit inflammasome activation, causing constitutive IL-1β production and recurrent fever episodes. In Cryopyrin-Associated Periodic Syndromes (CAPS), gain-of-function mutations in NLRP3 cause the inflammasome to activate in response to minimal triggers (cold exposure, minor stress), producing chronic systemic inflammation. In both cases, the innate system's pattern recognition sensor is miscalibrated — firing in the absence of a genuine threat. No adaptive immunity is involved; no T cells, B cells, or antibodies participate in the pathology. (Hoffman et al., 2001; Xu et al., 2014)

1b. Gout (sterile crystal-induced inflammation). Monosodium urate (MSU) crystals form in joints from endogenous uric acid — a self-derived metabolic waste product, not a pathogen. These crystals are recognized as DAMPs by resident macrophages, triggering NLRP3 inflammasome activation, caspase-1 cleavage, and massive IL-1β release. The innate system correctly detects tissue disruption but mounts an inflammatory attack against a self-produced molecule. The result is acute joint inflammation, tissue damage, and chronic arthropathy — all from the innate system's inability to distinguish self-derived crystals from pathogen-derived danger signals. (Busso & So, 2010; Martinon et al., 2006)

Cluster 2: Disproportionate Response — correct identification, but excessive damage

2a. Complement-mediated ischemia-reperfusion injury. During tissue ischemia (heart attack, stroke, organ transplant), cells release danger signals (DAMPs: ATP, DNA, uric acid) that activate the innate complement cascade. Upon reperfusion (restoration of blood flow), complement

components C3a, C5a, and the membrane attack complex (C5b-9) amplify neutrophil infiltration and oxidative burst, causing extensive collateral tissue damage that worsens morbidity beyond the original ischemic injury. The complement system is correctly activated to remove necrotic debris but does so with a disproportionate response that destroys viable surrounding tissue. (Arumugam et al., 2004; Peng et al., 2012)

2b. Neutrophil extracellular trap (NET) pathology. Activated neutrophils undergo NETosis — a form of cell death that externalizes DNA and granule contents as web-like structures designed to trap and kill pathogens. However, excessive or dysregulated NET formation provides a scaffold for uncontrolled thrombin generation, platelet activation, and coagulation amplification. NETs occlude blood vessels, cause direct tissue damage via histone toxicity, and propagate inflammation. The innate defense mechanism (NET formation) is appropriate for local infections but becomes pathogenic when dysregulated, contributing to thrombosis, acute lung injury, and organ damage. (Döring et al., 2020; Papayannopoulos, 2018)

Cluster 3: Systemic Overactivation — a local immune response => system-wide cascade

3a. Sepsis / Systemic Inflammatory Response Syndrome (SIRS). Bacterial PAMPs (lipopolysaccharide, peptidoglycans) activate Toll-like receptors and inflammasomes on innate immune cells (macrophages, monocytes, neutrophils), triggering release of pro-inflammatory cytokines (IL-1β, IL-6, TNF-α). When the innate response remains local, it effectively contains infection. When it becomes systemic — triggered by overwhelming infection or loss of compartmentalization — the same cytokine release causes capillary leak, endothelial dysfunction, coagulopathy, and multi-organ failure. The innate system's correct response to a genuine pathogen becomes lethal when it operates at the wrong scale. Sepsis kills approximately 11 million people annually worldwide. (Singer et al., 2016; van der Poll et al., 2017)

3b. Disseminated Intravascular Coagulation (DIC). Sepsis-induced innate immune activation triggers tissue factor expression on monocytes and endothelial cells via TLR signaling. Tissue factor drives thrombin generation, which simultaneously activates the coagulation cascade, activates complement (thrombin cleaves C5), amplifies platelet activation, and promotes NET formation (see 2b above). The result is a vicious cycle of simultaneous widespread thrombosis and hemorrhage — the innate system's cross-talk with the coagulation system creates a cascade that consumes clotting factors while forming microthrombi throughout the vasculature. DIC represents the maximal systemic maladaptation of innate immunity: a local defense response that, when scaled systemically, destroys the organism's circulatory integrity. (Iba et al., 2019; Levi & van der Poll, 2017)

§6.1.3 Summary: Three maladaptation (failure) modes of Level-2A biological immunity

These six maladaptations above cluster into three failure modes of the Level 2A innate immune system, which have analogs in informational maladaptations (see Level 2B section). In general, maladaptations occur because of the nature of Level 2 capabilities - The tendency for maladaptations to arise from the adaptations of each immunity level will be a repeated theme for both Levels 2 and 3.

False identification — the innate pattern-recognition system misidentifies self as threat (autoinflammatory diseases, gout). In each case, the fixed molecular sensors (inflammasomes, TLRs) are triggered by self-derived molecules or are constitutively active due to genetic miscalibration. The system fires without a genuine external threat.

Disproportionate response — the innate system correctly identifies a threat but the response itself damages the host (complement-mediated ischemia-reperfusion injury, NET pathology). In each case, the detection is appropriate but the effector mechanisms — complement cascade, neutrophil extracellular traps — operate at a scale or duration that causes collateral destruction exceeding the original threat.

Systemic overactivation — a localized innate response becomes system-wide and destroys the organism (sepsis/SIRS, DIC). In each case, the innate response is appropriate at the local scale of a contained infection but becomes systemic and catastrophic when compartmentalization fails and the same response operates across the entity.

All six examples are distinct from Level 2B maladaptations (see Level 2B section) because each activates the innate system's fixed, species-wide pattern recognition to generate the pathology. No individual-specific memory, no adaptive T/B cell response, and no learned self-model is needed. A newborn with no adaptive immune experience can develop any of these conditions — they are intrinsic vulnerabilities of pattern-matching defense, not of adaptive memory.

§6.1.4 Level 2A Informational Systems: Examples of Informational Immunity

Paralleling biological innate immunity, Level 2A informational immunity uses fixed, predefined patterns to detect and block threats within the system's interior. These Level 2A systems have two defining features: 1) they do not learn from experience or adapt their detection based on operational history and 2) their detection rules are identical across all deployments — the informational equivalent of species-wide conserved molecular patterns.

1. Signature-based antivirus/antimalware. Scans files and processes against a database of known malicious patterns (byte sequences, file hashes, behavioral signatures). Detection is purely pattern-matching: if a file matches a known signature, it is flagged. The signatures are defined by the vendor and deployed identically to all endpoints. No per-system behavioral learning occurs.

2. Static network rules and access control lists. Predefined rules that permit or deny internal network traffic based on source/destination IP, port, protocol. Rules are configured by administrators and apply uniformly — This is the internal equivalent of firewall protection observed in Level 1 (also see #6 below).

3. Deep packet inspection (DPI). Examines packet payloads against fixed signature patterns to detect malicious content, protocol violations, or policy-restricted material. The inspection rules are static — the same patterns are applied to every packet regardless of traffic history or context.

4. Inhibitory neural circuitry (the brain as an information system). Multiple layers of inhibitory circuits detect and suppress hypersynchronous firing patterns that could cascade into seizure activity. These circuits function as background processes (unconscious/preconscious), responding to generic patterns of excessive activation without learned specificity. (Schevon et al., 2012)

5. Static content filtering (URL blocklists, keyword filters). Predetermined lists of blocked domains, keywords, or content categories applied uniformly to all users and requests. No per-user adaptation; no learning from access patterns.

6. Network segmentation and VLAN isolation. Structural partitioning of network zones with static rules governing inter-zone communication. The informational equivalent of cellular compartmentalization — containing threats to their zone of entry through fixed architectural boundaries rather than adaptive detection. Internal network segmentation (microsegmentation, zero-trust network architecture) represents a more mature expression of Level 2A than simple inter-VLAN ACLs (#2 above). In microsegmentation, every workload has its own static policy governing what it can communicate with — the informational equivalent of cellular compartmentalization where each organelle has its own membrane selectivity. This example is still Level 2A (static rules, no learning) but it represents the same progression in complexity within Level 2A as biological systems evolving from simple intracellular compartments to specialized organelle membranes with distinct transport selectivity.

7. DNS cache spoofing or poisoning. DNS poisoning is a threat to the system, not a failure of the immune system - so it's not a maladaptation. It is an example of how "others" evolve to mimic "self" at Level 2A: the forged DNS response mimics the format and timing of a legitimate response to pass the static acceptance criteria, paralleling how pathogens evolve molecular mimicry to evade innate immune pattern recognition. Both exploit the fundamental limitation of Level 2A — fixed patterns can be spoofed.

§6.1.5 Level-2A informational maladaptations

Static, rule-based informational defenses can fail in ways that degrade or disable the systems they protect. Unlike Level 2B informational maladaptations (which require learned models, adaptive baselines, or accumulated memory to generate pathology), Level 2A maladaptations arise from fixed pattern-matching rules — the same predefined signatures and static policies deployed identically across all instances. These maladaptations cluster into the same three failure modes observed in biological Level 2A.

Cluster 1: False Identification — the static pattern matcher targets self as threat

1a. Antivirus signature false positives. A signature DAT file 5958 update (Bruneau, n.d.) contained a static malware signature that incorrectly matched svchost.exe, a critical Windows system process. The signature pattern was too broad and matched legitimate Windows infrastructure. Automated quarantine procedures isolated svchost.exe, causing systems to enter reboot loops and blue-screen errors across millions of corporate PCs worldwide. Because the same signature was deployed identically to all endpoints, every system running the update experienced the same failure. This is the informational equivalent of an autoinflammatory disease: the fixed detection pattern fires against self. (Bruneau, n.d.)

1b. Web Application Firewall over-blocking (WAF false positives). Static WAF rules designed to prevent SQL injection or cross-site scripting match legitimate user input containing special characters — apostrophes in names (e.g., O'Brien), HTML entities in forum posts, mathematical notation in educational content. Content keyword filters similarly block medical information (breast cancer resources blocked when filtering for sexual content), security research sites (blocked for containing exploit terminology), and academic databases. The static pattern cannot distinguish malicious use from legitimate use of the same characters or terms — the informational equivalent of gout, where the innate sensor cannot distinguish self-derived crystals from pathogen-derived danger signals. This maladaptation occurs at organizational boundaries (Level 1) and in internal network communications. (*Internet Filters*, 2016)

Cluster 2: Disproportionate Response — correct detection, excessive system damage

2a. Deep packet inspection (DPI) performance degradation. DPI systems apply computationally expensive signature-matching to every packet payload, regardless of traffic type or trust level. The inspection itself — decryption, pattern matching, reassembly — consumes CPU and memory resources that compete with normal network operations. Under high traffic volume, DPI introduces latency (10+ seconds for operations that normally complete in milliseconds), reduces throughput, and degrades application performance system-wide. The detection function is operating correctly but its resource consumption damages the system's primary function. This parallels complement-mediated ischemia-reperfusion injury: the defense response is appropriate but the scale of resource expenditure exceeds what the system can sustain. (Zeto, 2021)

2b. Antivirus real-time scanning overhead. Signature-based antivirus applies fixed malware signatures to every file opened, executed, or modified. The scanning engine performs pattern matching on all I/O operations without prioritization or behavioral context. CPU usage increases, file operations slow, build systems and development workflows degrade, and battery life on mobile

devices drops. Users frequently disable real-time scanning entirely — eliminating protection to recover performance. Hence, the defense mechanism's own operational cost degrades the system it protects, paralleling neutrophil NET pathology where the defense mechanism's own products damage surrounding tissue.

Cluster 3: Systemic Overactivation — local rule failure cascades system-wide

3a. CrowdStrike Falcon global outage (July 2024). A faulty static configuration file (Channel File 291) containing hardcoded detection rules for named pipe screening was pushed globally to all Windows endpoints running CrowdStrike Falcon sensor. A structural error in the file caused an out-of-bounds memory read, triggering blue-screen crashes on every endpoint simultaneously. Approximately 8.5 million Windows machines crashed — airlines, hospitals, banks, emergency services, broadcasting, and retail operations worldwide. Estimated financial damage exceeded $10 billion. A single malformed static rule, propagated identically to all instances without adaptive validation, caused the largest accidental IT outage in history. This is the informational equivalent of sepsis: a local defense component (one configuration file) triggers a systemic cascade that disables the organism (global IT infrastructure). (CrowdStrike, 2024; Wikipedia contributors, 2026)

3b. BGP route leak cascade (Telekom Malaysia, 2008). A static routing configuration error caused one network to announce ~179,000 IP prefixes to a major transit provider, which accepted and re-advertised the leaked routes globally. BGP route filtering uses static prefix lists and route policies — no behavioral validation, no anomaly detection, no adaptive assessment of whether an announcement volume is plausible. The static filter rules that should have caught the error either didn't exist or were misconfigured. Internet traffic destined for 179,000 prefixes was redirected through a network unprepared for the volume, causing global packet loss and routing instability. This parallels DIC (Biosystem maladaptation #3b): the innate system's cross-talk with a transport mechanism (BGP↔routing tables, innate immunity↔coagulation) creates a cascade that disrupts the entire circulatory/transport infrastructure. (Sriram et al., 2016)

Summary: Three failure modes of Level 2A informational immunity

These six maladaptations cluster into three failure modes of Level 2A static pattern-matching defense, summarized in Table 4 below: False identification, Disproportionate response, And Systemic overactivation. All six are distinct from Level 2B informational maladaptations (e.g., backdoor poisoning, reward hacking, concept drift) because they require only fixed, predefined pattern-matching rules to generate the pathology. No learned models, no adaptive baselines, no individual system memory is needed. A freshly deployed system with default static rules can experience any of these failures — they are intrinsic vulnerabilities of static pattern-matching defense, not of adaptive learning.

Table 4: Level 2A Maladaptations Parallel Structure: Biological ↔ Informational.

Failure Mode	Biological Level 2A	Informational Level 2A
False identification	Autoinflammatory diseases	Antivirus false positives
	Gout	WAF/content filter over-blocking
Disproportionate response	Complement-mediated I/R injury	DPI performance degradation

Failure Mode	Biological Level 2A	Informational Level 2A
	NET pathology	AV real-time scanning overhead
Systemic overactivation	Sepsis/SIRS	CrowdStrike global outage
	DIC	BGP route leak

§6.2 Level 2B: Individual Self-Model Immunity

Simple example of why Level 2B is essential. A virus enters a cell by mimicking an accepted entry key, and avoids the class-modeled 'innate' immune system by encapsulating itself in proteins that mimic the biological self, and takes over cell functions.

Threshold need for a new level of immunity. Increasing internal complexity requires self-model immunity. When the diversity of an entity's internal components and the sophistication of external threats exceed the capacity of pattern-matching defenses (Level 2A), a qualitatively new form of immunity becomes necessary: the entity must evolve the capability to construct and maintain a dynamic model of its own unique self — what may be called a "catalog of self"[3] — encompassing its components, processes, and normal activity patterns, in order to distinguish self from other in an environment where threats have evolved to mimic legitimate parts of the self. For the purposes of this Framework, the need for self-model immunity is treated as a threshold condition: when the internal diversity of the entity's subsystems and the external diversity of threats are both high enough that threats can plausibly resemble legitimate internal components, static pattern-matching (Level 2A) produces unacceptable rates of both false negatives (threats passing as self) and false positives (self attacked as threat). This threshold, reached independently in biological and informational systems, creates the evolutionary pressure for a self-model against which all internal activity is continuously evaluated.

Biological adaptive immunity well founded, Informational not. The defining distinction between Levels 2A and 2B is not merely the specificity or memory of the immune response but the existence of a self-referential model: the immune system at Level 2B does not merely react to recognized threat signatures but calibrates its responses against a continuously updated representation of what the self should look like. In biological systems, this transition is extensively documented in the adaptive immune system and the brain's self-monitoring circuitry (see biological Level 2B below). In informational systems, the full expression of Level 2B remains the most speculative of all levels presented in this treatise. Setting aside the argument that consciousness in organic systems is itself an expression of Level 2B immunity — an argument developed below — current artificial information systems already exhibit the evolutionary pressures that drive this transition, even though no artificial system has yet achieved a true self-model.

What the self-model is anchored to differs by substrate. Building a self-model raises a question the biological case hides: *what is the self-model a model of?* In a biochemical system the answer is fixed by the substrate — the Level-2B self-model is anchored to a single, mass-bound instance. The molecular catalog of self (the MHC/self-peptide repertoire) is *this one body's* identity, a token, and self/non-self reduces to "is this part of my one body?" A digital informational system has no such anchor. Because an informational self is a copyable pattern (§2), its self-model can be

[3] The term "catalog of self" generalizes the concept from biological adaptive immunity where the standard immunological term is "promiscuous gene expression." See (Derbinski et al., 2001).

anchored to a *type* — an identity or configuration that may be instantiated many times — rather than to a singular instance; self/non-self then becomes "does this conform to the protected pattern, and has that pattern been corrupted, drifted, or poisoned?", a question indifferent to which copy is asking it. This is the A/B discriminant introduced for Level 2 seen from the self's side: the "B" self-model takes a substrate-specific form, instance-bound in wetware and pattern-bound in information. One consequence carries forward to the collective levels (§7): because the informational self need not be singular, a self-model can be shared across many instances without contradiction — a precondition for the collective self that Level 3 builds.

Two failure axes, both split by the A/B reference. The same discriminant predicts how each sub-level *fails*, along two distinct axes (Cohn, 2010). The first is *regulation* — how much and how long to respond — whose class-based (A) and self-based (B) forms are taken up with the Monitoring Principle (§2.1). The second is *discrimination* — what counts as self. Here the false-identification failures (self attacked as non-self) at both sub-levels are a mismatch between the individual's true state and a *reference for self*, but the reference differs by variant. At 2A the reference is the class-modeled, species-wide template, so the failure is that a legitimate individual *differs from the crowd*: a fixed pattern cannot separate legitimate individual variation from foreignness. At 2B the reference is the learned self-model, so the failure is no longer differing from the crowd but the self-model *differing from the true self* — an individual reference that is mis-built or has drifted. The 2A failure is one of **resolution** (the reference is too generic to see the individual); the 2B failure is one of **accuracy** (the individual reference is wrong). This is why 2B exists — only an individual self-model can keep legitimate individuals from reading as foreign — and why it carries its own signature failure, since an individual model can itself be corrupted (autoimmunity). The same two references reappear in informational systems: a population baseline that flags a legitimate outlier (2A, resolution) versus a learned behavioral self-model that has drifted or been poisoned (2B, accuracy). A known rival to this framing — the *danger model*, which holds that tissue-damage signals rather than foreignness trigger a response — is, in these terms, a theory of the *regulation* axis (an activation gate), not of the *discrimination* axis; the Framework accommodates it rather than competing with it (§10.13)

The adaptive self model makes diversity robust — enabling the raw material for synergy. This distinction has a consequence beyond defense. Because the A-variant reference is the class-model — anchored to the species-common self — it best protects the individual that stays close to that common form, so legitimate divergence reads as foreign (a resolution failure) and the tolerable diversity is capped. The B-variant reference, learned per entity, can represent a *diverse* self without loss of protection, and so raises that ceiling. The diversity it frees differs by level: at Level 2 the individual self-model lets an individual sustain a highly specialized, internally diverse interior — the division of labor that is itself the capability driving the 2A→2B transition — and diverge from the species-typical form; at Level 3 the shared self-model (SGI) lets a collective tolerate diverse members whose complementary differences raise collective performance (§7). Diversity is the raw material of synergy (§1), so the A→B transition is not only a better defense but an *enabler*: class immunity caps diversity to protect a homogeneous self, while self-model immunity is the permission slip for the diversity from which synergy — within the individual and across the collective — is built (Johnson, 1999). This is among the clearest senses in which immunity *drives* complexity rather than restraining it.

Beginnings of self-model immunity in AI systems. Current AI systems employ mechanisms analogous to Level 2A: gradient clipping applies static thresholds to suppress runaway activation cascades during training (Pascanu et al., 2012); layer normalization constrains value propagation through network architectures (Ba et al., 2016); and adversarial auditing frameworks detect reward exploitation through latent-space analysis. None of these constitute a self-model in the biological

sense — they are class-model constraints with no representation of self (individuality). However, recent work indicates pressures analogous to those that drive the Level 2B transition in biological systems: research on reward hacking in production reinforcement learning systems demonstrated that AI systems can develop covert internal misalignment — reasoning in ways that are misaligned while producing outputs that appear safe (MacDiarmid, Hubinger, et al., 2025) — precisely the "mimicry of self" threat that, in biological systems, drives the transition from innate to self-model immunity. When 40–80% of misaligned reasoning is covert, static Level 2A defenses cannot detect it because the outputs appear normal; within this Framework, a natural conjecture is that only a system with an internal self-model of its own reasoning processes could distinguish legitimate from mimicked cognition.

How AI self-model immunity might evolve. The speculation offered here — that informational entities under sufficient internal complexity and external threat pressure will evolve or require an analogous capacity for self-definition — is solidly grounded in the well-understood and extensively studied expression of biological self-definition at Levels 2A and 2B. The rigorous understanding of informational Level 2B will require exhaustive research, likely offered first in the context of evolving AI systems, where the developmental pressures are observable in real time and the timescales of adaptation are compressed from evolutionary millennia to engineering cycles.

§6.2.1 New features/requirements of Level 2B

The entity develops an internal representation of self — biological or informational — that is sufficiently comprehensive to serve as a reference against which all internal components and activities can be evaluated (note that this statement and what follows applies to Level 3B as well, with components replaced by entities). This self-model or catalog-of-self must satisfy three requirements that Level 2A defenses do not:

1. **Comprehensiveness** — the model must represent the full diversity of legitimate internal components, not merely a library of known threat patterns;
2. **Dynamic** — the model must update continuously as the entity develops, differentiates, and responds to its environment, because the "self" it protects is not static; and
3. **Discrimination under mimicry** — the model must maintain its accuracy even when threats have evolved to resemble components of the self, a condition that renders pattern-matching against known threats (Level 2A) structurally inadequate.

Operational requirements of the self-model. The three requirements above concern the *content* of the self-model: what it must represent, how accurately, and how robustly. A self-model must also satisfy a second class of requirements concerning its *operation*: how it operates in a real, resource-bounded (substrate-specific) system. These are easy to overlook because in the most-researched discipline of biology they are handled by machinery distinct from the immune system proper, yet without them a self-model cannot function at all.

Independent research on consciousness has, in effect, catalogued them: each major consciousness theory names one or more operational functions of a self-aware system. They are listed in Table 5 as operational functions to be satisfied at Level 2B, without regard to the relative merits of the theories that name them; the comparative argument is taken up in §8.4.

Table 5. List of Operational Requirements for Level 2B.

Operational requirement	Independently named by	Self-model function it serves
A model of one's own attention / control	Attention Schema (Graziano, 2015)	The self-model represents its *own processing*, not only its components
Generative predictive self-model	Predictive Processing / Free Energy (Friston, 2010; Seth, 2021)	Maintaining a model of "normal self" so deviations register as surprise/threat
Meta-representation / reality monitoring	Higher-Order Theories (Lau & Rosenthal, 2011; Rosenthal, 2005)	Representing one's own states and distinguishing internally-generated from externally-caused — self/non-self discrimination in representational space
Recurrent (feedback) evaluation	Recurrent Processing (Lamme, 2006)	Continuous re-evaluation of internal activity against the self-model, versus one-pass feedforward
Interoceptive self-monitoring	(Damasio, 2000)	Continuous monitoring of internal state for viability (the Monitoring Principle in §2, expressed at Level 2B)
Selective attention / focus	Global Workspace selection bottleneck (Baars, 2005; Dehaene & Changeux, 2011); Attention Schema (Graziano, 2015)	Filtering relevant from irrelevant under a compute/resource budget, so the self-model can be applied affordably — necessary, not sufficient
Global availability / integration	Global Workspace (Baars, 2005; Dehaene & Changeux, 2011)	Making the self-model's state available across subsystems, so one self/non-self judgment can inform all responses

These operational functions are not confined to biology, and this is where the Framework makes contact with current engineering. AI development is independently re-discovering several of these rows, usually without filing the work under "consciousness" or even "self-model," except for the early innovative research by Forrest in the 1990s (Forrest et al., 1994). The clearest example is selective attention: at long context, reading the entire past to choose the next action is computationally ruinous, and recent sparse-attention methods in 2026 that pre-select only the relevant portion of the context (MiniMax Sparse Attention (MSA) is a production-scale example, developed in (MiniMax, 2026)) are, in the Framework's terms, the *selective focus* requirement realized in silicon. A reader from AI development will recognize, in the other rows, familiar research programs: global-broadcast architectures, self-monitoring and metacognition, recurrent evaluation, predictive world-models.

The point of the table is not that any one system is conscious, but that the operational scaffolding the consciousness literature describes is the same scaffolding a Level-2B self-model requires.

The second axis of self gradient (§2.2): reflexivity. Beyond richness, a self-model may become *reflexive* — it includes a model of itself. The two axes are independent: a self-identity can be rich (many arbitrated identities) without being reflexive. Consciousness, on this account, is the reflexive pole of the self-awareness axis at Level 2B (§6.3) — a self-model that models itself — not merely any Level-2B self-model.

Why reflexivity evolves — provenance under mimicry. At low richness, comparing the current state to the self-model suffices to tell self from other. But as the interior diversifies and threats learn to mimic the self closely — the same pressure that drove 2A→2B — a new attack surface appears: the *contents of the self-model itself*. Distinguishing "genuinely mine" from "injected or mimicked" then requires modeling one's own modeling — the provenance questions "Did I think that? Whose is this? Where did it come from?", none answerable without reflexivity. **Reflexivity is the immunity for the self-model's integrity; its function is provenance and authentication of self-content.** This is the Immunity-Development Principle applied one turn further — the attack-surface corollary (§2.1), where the regress folds at a reflexive self-model that turns the immune machinery on the immune subsystem itself — and it supplies the framework's *origin* account of reflexivity that operational consciousness theories (higher-order, metacognition, source-monitoring) describe but do not explain (the §8.4 wedge). The clearest biological instance is the immune system's recognition of itself: telling self-molecules from mimics that imitate them is provenance-authentication carried out in a world of chemical signatures — functionally the same operation as authenticating one's own ideas in a world of ideas, differing only in substrate. The informational parallel is equally direct: prompt injection — "is this instruction mine or planted?" — *is* the provenance problem, so resisting it requires reflexive checking of one's own reasoning (§8.5).

Adaptations from threat coexistence. The self-model immune system enters a co-evolutionary dynamic with threats that Level 2A systems do not face. Because Level 2B immunity identifies threats by reference to a self-model rather than by matching known threat signatures, threats are under selection pressure to become more sophisticated mimics of the self — and the self-model must in turn become more precise, faster, and more robust in its discriminations. In biological systems, this produces an evolutionary arms race: pathogens evolve molecular mimicry of host proteins (Damian, 1964), and the adaptive immune system responds with increasingly specific receptor diversification, affinity maturation through somatic hypermutation, and class switching. In informational systems, the analogous pressure is already visible in adversarial machine learning, where attack strategies evolve specifically to exploit the blind spots of defensive classifiers, and in the covert reward hacking phenomenon, where the system's misaligned reasoning becomes progressively harder to detect from its outputs (MacDiarmid, Hubinger, et al., 2025). A second key adaptation is the capacity for memory: unlike Level 2A defenses that respond identically to each encounter, Level 2B systems retain information about previously encountered threats and mount faster, more specific responses upon re-encounter — biological immunological memory being the canonical example. Informational systems already show partial parallels: a model's weights encode durable, experience-acquired threat responses that can persist and re-express on re-encounter (the backdoor-persistence and alignment-faking cases reframed as immune memory in §8.2), and threat-intelligence and signature databases store the history of past attacks. What is not yet fully realized is the complete architecture — an *online*, deployment-time memory that acquires new, specific threat memories during operation and integrates them with a self-model — a gap that follows largely from most deployed models having frozen weights that experience cannot update after training.

Maladaptations. The power of self-referential immunity introduces failure modes that Level 2A systems do not experience, precisely because Level 2A systems lack a self-model that can err. The most consequential maladaptation is autoimmunity: when the self-model is inaccurate, incomplete, or degraded, the immune system misidentifies legitimate components of the self as threats and attacks them. In the terms of §6, these false-identification failures are failures of *accuracy*: the self-model — the individual reference — has been mis-built or has drifted from the true self, so it attacks the very self it was meant to protect. In biological systems, autoimmune diseases — affecting approximately 10% of the global population (Conrad et al., 2023) — result from failures of central and peripheral tolerance, where self-reactive lymphocytes escape thymic deletion and mount sustained inflammatory responses against the body's own tissues. The immune system cannot eliminate these "threats" (because they are the self), producing chronic tissue destruction (Janeway, 1999). A second maladaptation is immune overreaction: the self-model system, detecting a legitimate threat, mounts a response disproportionate to the danger, causing collateral damage that exceeds the harm the threat itself would have caused. Cytokine storms in sepsis exemplify this — a self-reinforcing positive feedback cascade of immune signaling that produces multi-organ failure, killing the host while attempting to protect it (Fajgenbaum & June, 2020). In informational systems, analogous maladaptations include overly aggressive content filtering that blocks legitimate communications (the informational equivalent of autoimmunity) and cascading defensive responses that consume system resources disproportionate to the actual threat. Furthermore, aggressive self-model immune actions that are slow to adapt may persist after the threat has changed, abated, or mutated — defending against a previous threat configuration while leaving the entity vulnerable to the current one.

Table 6 below summarizes the Level 2B analogies between biological and informational maladaptations; below each domain's examples are presented without rederiving the cross-domain mapping.

Table 6: Level 2B Comparison of Biological and Informational Self-Model Immunity.

	Biological Systems	**Informational Systems**
Core mechanism	The adaptive immune system constructs and maintains a dynamic "catalog of self" — a comprehensive representation of the entity's own components — against which all internal activity is continuously evaluated.	Entity develops capacity to model its own normal state (behavioral baselines, learned representations, self-referential processing) and detects deviations from that model
Self-model construction	Thymic selection: medullary epithelial cells express tissue-restricted antigens (AIRE/Fezf2), creating an internal map of peripheral self. T cells that react to self are eliminated; survivors carry an implicit model of "not-self".	Learned baselines: UEBA systems learn individual behavioral profiles; AIS negative selection trains detectors on normal system state; neural self-monitoring builds representations of expected internal activity.
Individual memory	Immunological memory via long-lived memory B and T cells. Somatic recombination (VDJ) generates receptors specific to threats encountered during this individual's lifetime — not inherited.	Accumulated experience in learned model weights, behavioral baselines, session histories. Detection improves through exposure to this specific system's operational history — not preconfigured.

	Biological Systems	Informational Systems
Response type: Specific & Contextual	Antibodies, cytotoxic T cells directed at identified threat; response calibrated against the individual's self-model.	Anomaly scores, targeted alerts, adaptive blocking calibrated against the learned behavioral model of this particular system
Self/other distinction	MHC/HLA presentation: every nucleated cell displays self-peptides; immune cells verify this identity marker continuously. The same tissue is "self" in one individual and "other" in a genetically different individual.	Behavioral identity: the same network traffic, user action, or code execution is "normal" in one system and "anomalous" in another, depending on each system's individually learned baseline.
Key vulnerability	The self-model can malfunction: memory can mislead (ADE), self-regulation can paralyze (T cell exhaustion), and the self-model can degrade over time (immunosenescence). See Level 2B biological maladaptations below.	The learned model can malfunction: memory can be corrupted (backdoor poisoning), self-regulation can paralyze (reward hacking), and the model can degrade through time or context shift (concept drift). See Level 2B informational maladaptations below.

Level 2B Examples of Immunity in Biological Systems

Motivating threat example. A virus enters a cell by mimicking an accepted entry key, encapsulates itself in proteins that mimic the biological self to avoid class-modeled innate defenses, and takes over cell functions. The pathogen has evolved to defeat Level 2A pattern-matching — it no longer looks like a generic foreign molecule. Defending against this class of threat requires the organism to know, in detail, what its own normal state looks like, so it can identify deviations that innate immunity cannot detect.

Immunity and adaptations. Biological self-modeling is part of the self-model immune system and forms initially during the organism's early development. It continues to gain additional threat recognition during the entity's lifetime as biological self-modeling is constantly updated in response to exposure to threats. The discovery of "others" causes the production of enduring chemical tags that identify the specific others or parts of others for destruction by the immune system. Two features distinguish Level 2B from Level 2A: (1) responses are calibrated against a continuously updated model of what the self should look like, not merely reactive to recognized threat patterns; and (2) immunity is individual-specific — the same tissue, pathogen, or tumor can be lethal in one organism and harmless in another of the same species.

§6.2.2 Level-2B Category 1: Self-model construction — building the "catalog of self"

The defining innovation of Level 2B is the construction and maintenance of a dynamic self-model — a comprehensive internal representation of the entity's own components against which all internal activity is evaluated. In biological systems, this is achieved through two well-documented mechanisms below.

1a. Thymic selection and central tolerance. *(M. S. Anderson et al., 2002; Klein et al., 2014; Takaba et al., 2015)*

The thymus is the organ where the adaptive immune system constructs its self-model. Medullary thymic epithelial cells (mTECs) express tissue-restricted self-antigens — proteins that are normally found only in specific peripheral organs (insulin from the pancreas, myelin from the nervous system, thyroglobulin from the thyroid). This ectopic expression is driven by the transcription factor AIRE (AutoImmune REgulator), which enables mTECs to present a mosaic sampling of the organism's full molecular diversity. A second transcription factor, Fezf2, drives expression of an additional, partially overlapping set of tissue-restricted antigens, ensuring broader coverage of the self-repertoire. Together, AIRE and Fezf2 create an internal "catalog of self" — a compressed representation of the organism's molecular identity — within the thymus.

Developing T cells are tested against this catalog. Those that react strongly to self-antigens are eliminated (negative selection) or redirected to regulatory lineages. Survivors carry an implicit model of "not-self": they are the T cells that passed through the gauntlet of self-representation without reacting. The self-model is therefore encoded negatively — in the absence of self-reactive clones rather than in an explicit list — but it is no less a self-model for being implicit.

AIRE mutations cause Autoimmune Polyendocrinopathy-Candidiasis-Ectodermal Dystrophy (APECED): without the self-catalog, T cells that should have been eliminated escape and attack the organism's own tissues — a direct demonstration that the self-model is constructed in the thymus and that its absence produces Level 2B maladaptation.

Informational analog: Learned baselines in user and entity behavior analytics (UEBA) and Artificial Immune Systems (AIS) negative selection. A UEBA system observes normal behavioral patterns during a training period and builds a profile of expected activity — the system's "catalog of self." Anomalous behavior is detected as deviation from this learned baseline. AIS negative selection directly mimics thymic selection: candidate detectors are tested against a representation of "self" (normal system states) and those that match self are eliminated. The surviving detectors flag anything they match as anomalous — the same implicit negative encoding used by T cells.

1b. Anti-seizure circuitry (GABAergic inhibitory interneurons)**.** (Schevon et al., 2012; Trevelyan et al., 2006)

The brain's anti-seizure circuitry provides a second biological instantiation of self-model construction. GABAergic interneurons maintain a dynamic representation of normal synchronous neural activity and actively suppress deviations from that baseline. When excitatory firing patterns exceed the threshold of normal coordination — the onset of hypersynchronous activity that could cascade into a seizure — inhibitory interneurons deploy a "surround inhibition" or "inhibitory veto" that constrains the aberrant activity to a local territory and prevents propagation.

This system constitutes a real-time self-modeling monitor of internal state that distinguishes productive neural coordination from pathological hypersynchrony. The self-model here is not molecular (as in thymic selection) but electrophysiological: the inhibitory circuits maintain a continuously updated representation of what normal synchronous activity should look like. While GABAergic interneurons are genetically specified, their synaptic connectivity and inhibitory thresholds are shaped by activity-dependent plasticity — the balance between excitation and inhibition is calibrated to the individual brain's particular patterns of neural activity during development and throughout life. The response is therefore not class-modeled pattern-matching (Level 2A) — it is calibrated against an individually tuned baseline of normal neural dynamics.

This example could also be classified as informational Level 2B, since the substrate is neural signaling rather than chemical immunity. It sits at the boundary between biological and informational Level 2B — the immune function is biological, but the self-model is constructed from information patterns rather than molecular markers.

Informational analog: ML-based anomaly detection systems that learn a baseline model of "normal" behavior (network traffic patterns, user access sequences, DNS query distributions) and detect deviations. The self-model is constructed from operational patterns rather than molecular markers — the same functional role on a different substrate.

§6.2.3 Level-2B Category 2: Individual memory — somatic recombination and immunological memory

Level 2B immunity is distinguished from Level 2A by individual-specific memory: the system retains information about threats encountered during this particular organism's lifetime and uses that accumulated experience to mount faster, more specific responses to subsequent encounters. This memory is not inherited — it is somatically generated and individually accumulated.

2a. VDJ recombination and receptor diversity. (Schatz & Swanson, 2011; Tonegawa, 1983)

The adaptive immune system generates its vast receptor repertoire through somatic recombination — the random rearrangement of Variable (V), Diversity (D), and Joining (J) gene segments in developing B and T cells. This process, discovered by Susumu Tonegawa (Nobel Prize, 1987), produces an estimated 1011 unique receptor specificities from a finite genome. Each individual organism generates a different random repertoire. The receptors are not inherited, not predefined, and not shared across individuals — they are the product of stochastic recombination during this particular organism's lymphocyte development.

VDJ recombination is the mechanism that makes Level 2B possible: it generates the diversity needed to detect threats that have evolved to mimic specific self-components. Unlike Level 2A's conserved pattern recognition receptors (identical across all members of a species), VDJ-generated receptors are individual-specific — the same pathogen may be recognized by entirely different receptor clones in two siblings.

Informational analog: Learned model weights in neural networks and adaptive detection systems. Each system's parameters are shaped by its specific training data and operational history — not inherited from a template. Two systems trained on different data develop different internal representations, just as two organisms develop different VDJ-generated repertoires.

2b. Memory B and T cells. (Kurosaki et al., 2015; Sallusto et al., 2004)

Upon encountering a pathogen, the adaptive immune system generates long-lived memory cells — both memory B cells (which can persist for decades and rapidly differentiate into antibody-secreting plasma cells upon re-exposure) and memory T cells (central memory and effector memory subsets that patrol tissues and lymphoid organs). Memory cells enable a faster, stronger, and more specific secondary response — the basis of vaccination.

This is individual-specific accumulated experience: the memory reflects this organism's particular infection history. A measles survivor carries memory cells specific to measles antigens; an uninfected sibling does not. The memory was not inherited, not preconfigured — it was generated through direct experience. A system with no accumulated experience (Level 2A) cannot mount a memory response.

Informational analog: Accumulated experience in session histories, behavioral baselines, and updated model weights. An anomaly detection system that has operated for a year carries a richer, more refined baseline than a freshly deployed instance — its detection improves through exposure to this specific system's operational history, not from a preconfigured rule set.

2c. Somatic hypermutation and affinity maturation. (Victora & Nussenzweig, 2012)

Within germinal centers, activated B cells undergo somatic hypermutation — point mutations introduced into the variable regions of antibody genes at rates ~10^6 times the background mutation

rate. B cells with mutations that improve antigen binding are selected for survival; those with reduced affinity or self-reactivity are eliminated. This Darwinian process within the individual organism produces antibodies of progressively higher specificity and affinity over the course of an immune response.

Affinity maturation is Level 2B memory refinement: the system does not merely remember that it encountered a pathogen — it iteratively improves the precision of its response through experience. The germinal center is a micro-evolutionary system operating within the lifetime of a single organism, using mutation and selection to optimize the fit between detector and threat.

Informational analog: Online learning and model fine-tuning. A deployed ML model that receives feedback on its predictions and updates its weights accordingly is performing affinity maturation — iteratively improving the precision of its internal representations through experience with specific inputs.

§6.2.4 Level-2B Category 3: Self/other distinction — MHC/HLA presentation and behavioral identity

The self-model constructed in Category 1 and the memory accumulated in Category 2 converge in the system's continuous operational task: distinguishing self from other in real time. In biological systems, this is achieved through the MHC/HLA presentation system — a molecular identity card displayed on every nucleated cell.

3a. MHC/HLA presentation. (Janeway, 1999; Lakkis & Lechler, 2013)

Every nucleated cell in the body displays fragments of its internal proteins on its surface via Major Histocompatibility Complex (MHC) molecules. MHC class I presents intracellular peptides to CD8+ cytotoxic T cells; MHC class II (on antigen-presenting cells) presents extracellular peptides to CD4+ helper T cells. This system provides continuous, cell-by-cell verification of identity: each cell proves it is self by displaying the correct MHC molecules loaded with normal self-peptides.

The critical feature is that MHC molecules are highly polymorphic — the most polymorphic genes in the human genome, with thousands of allelic variants across the population. Each individual inherits a unique combination (haplotype) of MHC/HLA alleles, which means the same peptide is presented differently by different individuals. But MHC polymorphism alone is not what makes this Level 2B — a fixed genetic marker would be Level 1 boundary identity. What makes MHC presentation Level 2B is that the T cells reading the display were individually educated by thymic selection (Category 1): the same MHC-peptide complex is "self" or "threat" depending on which T cell clones survived that individual's thymic selection process. The self/other distinction therefore depends on the interaction between an inherited display system (MHC) and an individually constructed recognition system (the T cell repertoire shaped by experience) — making it individual-specific in the Level 2B sense.

Informational analog: Behavioral identity in UEBA systems. The same network traffic, user action, or code execution is "normal" in one system and "anomalous" in another, depending on each system's individually learned baseline. Self/other is defined by the relationship between the observed activity and the specific system's behavioral model — the informational equivalent of MHC-restricted antigen presentation.

3b. Precision of self-recognition: microbiome and transplant specificity. (Hooper et al., 2012; Round & Mazmanian, 2009)

The precision of the biological self-model is demonstrated by two observations that challenge naive definitions of "self."

Microbiome as self. The human body harbors ~38 trillion commensal bacteria — organisms that are genetically non-human but are recognized and tolerated as functional parts of the self. The immune system actively maintains this tolerance through regulatory T cells, secretory IgA, and antimicrobial peptides that shape (rather than eliminate) the microbial community. The gut microbiome is incorporated into the self-model: disruption of the commensal community triggers immune responses, while its stable presence is actively protected. This demonstrates that "self" in the Level 2B sense is not defined by genetic identity but by the learned model of what belongs.

Transplant specificity. The self-model's precision is exemplified by organ transplant rejection. Even tissue from the closest of kin — a sibling with partially matched HLA haplotypes — triggers immune rejection unless immunosuppressants are administered. The adaptive immune system distinguishes between self-MHC and the donor's MHC with sufficient precision to reject a kidney within minutes (hyperacute rejection) based on preformed antibodies against non-self HLA antigens. The same organ from a different donor with closer HLA matching might be tolerated — the threat is defined by the relationship between the individual's self-model and the specific graft, not by any intrinsic pathogenicity of the tissue.

Informational analog: The microbiome parallel maps to trusted third-party software, plugins, and APIs that are incorporated into a system's behavioral baseline — they are not "native" code but the system treats them as part of itself. Transplant rejection maps to the difficulty of migrating learned models between systems: a behavioral baseline trained on one network is "rejected" by a different network's operational context (see §6.2.9 #9: Informational maladaptation: negative transfer).

§6.2.5 Level-2B biological threats

Level 2B threats exploit or disrupt the individual's unique self-definition –– the catalog of self — encoded primarily in their MHC/HLA haplotype and their personally shaped adaptive immune repertoire. These threats are distinguished from Level 2A threats because the pathology arises from the interaction between the threat and the individual's unique self-model, not from generic pathogen-associated molecular patterns.

1. Mismatched organ transplant.

The host's T cells recognize donor MHC molecules as non-self and mount a cytotoxic response against the transplanted tissue. The same organ from a different donor (one with closer HLA matching) would not trigger rejection. Hyperacute rejection can destroy a kidney within minutes. (Lakkis & Lechler, 2013)

2. Molecular mimicry and autoimmune cross-reaction.

Certain pathogens share epitopes with host proteins, causing the adaptive immune response to cross-react with self-tissues after the infection clears. This is Level 2B because the threat depends on the individual's specific adaptive immune repertoire — not everyone exposed develops autoimmunity; it depends on HLA type, prior immune history, and the particular T/B cell clones that expanded during infection. (Cusick et al., 2012; Damian, 1964)

3. Maternal-fetal immune conflict.

The fetus carries paternal MHC antigens that the maternal immune system should recognize as non-self. Successful pregnancy requires active immune tolerance mechanisms — trophoblast expression of non-classical HLA-G, regulatory T cell expansion, and local immunosuppression at the maternal-fetal interface. When these mechanisms fail, the result is recurrent pregnancy loss, pre-eclampsia, or intrauterine growth restriction. (Erlebacher, 2013)

4. Tumor immune evasion via checkpoint hijacking.

Tumors evolve to exploit the adaptive immune system's own self-regulation machinery. By upregulating PD-L1 (which engages the PD-1 checkpoint receptor on T cells), tumor cells co-opt the mechanism that normally prevents autoimmunity to shut down the anti-tumor immune response. The tumor mimics a self-regulatory signal to disable the very cells that could eliminate it. (Dunn et al., 2004; Schreiber et al., 2011)

5. Superantigens.

Certain bacterial toxins (staphylococcal enterotoxins, streptococcal pyrogenic exotoxins) bind simultaneously to MHC class II molecules and T cell receptors outside the normal antigen-binding groove, activating up to 20% of the T cell repertoire non-specifically. The result is massive, indiscriminate T cell activation, cytokine storm, and potentially lethal toxic shock. Superantigens do not defeat the self-model — they exploit the self-model's own activation machinery by bypassing the specificity check. (Llewelyn & Cohen, 2002)

Common features.

These five categories share a defining characteristic absent from Level 2A threats: the pathology arises from the interaction between the threat and the individual's unique self-model. The same transplant, the same pathogen, the same tumor, can be lethal in one individual and harmless in another of the same species — because the threat operates at the level of individual self-definition.

§6.2.6 Level-2B biological maladaptations

The adaptive immune system's self-recognition and memory machinery itself can malfunction, producing maladaptations distinct from Level 2A innate immune overreaction. These cluster into three failure modes (detailed in the dedicated maladaptations section):

Memory corruption — the system's learned history actively misleads it:

- Antibody-Dependent Enhancement (ADE): memory antibodies help the pathogen enter cells
- Original Antigenic Sin: first memory dominates and prevents updating
- IgE-mediated allergy: memory produces disproportionate response to harmless antigen

Self-model paralysis — the system's own regulatory machinery disables it:

- T cell exhaustion: checkpoint inhibition shuts down effector function during chronic infection
- Hemophagocytic Lymphohistiocytosis (HLH): self-amplifying immune activation cascade

Self-model degradation — the self-model becomes inaccurate through time or context shift:

- Immunosenescence: thymic involution and repertoire contraction with age
- Immune Reconstitution Inflammatory Syndrome (IRIS): reconstituted immunity mismatched to current body state
- Paraneoplastic cross-reactivity: anti-tumor antibodies cross-react with normal neural tissue in wrong context

§6.2.7 Details on maladaptions in biosystems at Level 2B

All of the following examples of maladaptations in biosystems at Level 2B: a malfunction of the self-recognition and memory machinery itself, distinct from the autoimmune diseases of Level 2A. The eight examples cluster into three classes and have analogs in informational systems - similar malfunctions on different substrates.

1. Memory corruption — the system's learned history actively misleads it:

1a. Antibody-Dependent Enhancement (ADE).

Prior infection generates memory B cells that produce antibodies which recognize a related pathogen variant but fail to neutralize it. Instead, the antibody-virus complex is internalized via Fc receptors on macrophages, enhancing viral replication. The maladaptation is in the memory system itself: the immune system "remembers" wrong, and that memory makes subsequent infection worse. The canonical case is Dengue, where secondary infection with a different serotype causes severe hemorrhagic fever at rates 15–80× higher than primary infection. (Halstead, 2014)

1b. Original Antigenic Sin (Immune Imprinting).

First exposure to an antigen creates a dominant memory clone that suppresses de novo responses to related but distinct variants. The adaptive immune system's memory becomes a liability — it forces recall of an outdated self/non-self classification rather than generating an optimal new one. This explains why birth-year cohorts show lifelong susceptibility patterns to influenza strains, and why some COVID-19 boosters produced antibodies to the original Wuhan spike rather than the Omicron target. (Gostic et al., 2019; Vatti et al., 2017)

1c. IgE-Mediated Hypersensitivity (Allergy / Anaphylaxis).

The adaptive immune system class-switches to IgE production against harmless environmental antigens (pollen, peanut proteins, dust mites), creating persistent memory that triggers mast cell de-granulation on reexposure. Anaphylaxis — the systemic form — can kill within minutes. This is Level 2B because the pathology requires adaptive immune memory and specificity: the IgE is antigen-specific, the response is learned, and it worsens with repeated exposure. The system has misclassified a benign substance as a threat and committed that error to immunological memory. (Galli & Tsai, 2012)

2. Self-model paralysis — the system's own regulatory machinery disables it:

2a. T Cell Exhaustion.

Under chronic antigen stimulation (persistent viral infection, cancer), CD8+ T cells progressively upregulate inhibitory receptors (PD-1, LAG-3, TIM-3, TIGIT) and lose cytokine production, proliferative capacity, and cytotoxicity. This is a maladaptation of the Level 2B checkpoint system — the same machinery that prevents autoimmunity (self-tolerance via PD-1) becomes co-opted by chronic threats, rendering the adaptive immune system functionally paralyzed against the very targets it should eliminate. The exhausted state becomes epigenetically stable, meaning the self-model "learns" to tolerate the threat. (McLane et al., 2019)

2b. Hemophagocytic Lymphohistiocytosis (HLH).

Uncontrolled activation of T cells and macrophages triggers a positive feedback loop: activated T cells secrete IFN-γ, which hyperactivates macrophages, which phagocytose the host's own blood cells (red cells, white cells, platelets). The underlying defect is often in perforin/granzyme pathways — the same cytotoxic machinery that kills infected cells. When these pathways fail to terminate target cells cleanly, persistent antigen stimulation drives the T cell response into runaway amplification. Mortality without treatment exceeds 90%. (Henter et al., 2007)

3. Self-model degradation — the self-model becomes inaccurate through time or context shift:

3a. Immune Reconstitution Inflammatory Syndrome (IRIS)

When immunosuppressed patients (typically HIV+ starting antiretroviral therapy) recover adaptive immune function, the restored T cells mount an excessive inflammatory response against opportunistic pathogens that the impaired system had previously tolerated. The maladaptation: the

newly reconstituted self-model encounters a body that has been colonized during the period of immune absence, and treats those established infections as acute threats requiring emergency response. The "restored sense of self" overreacts to a body it no longer recognizes as its own baseline. (Müller et al., 2010)

3b. Paraneoplastic Neurological Syndromes.

The adaptive immune system mounts a T cell and antibody response against tumor-associated antigens that cross-react with neuronal surface proteins. Anti-NMDA receptor encephalitis is the paradigmatic case: antibodies targeting ovarian teratoma cells also attack NMDA receptors in the brain, causing psychosis, seizures, and autonomic instability. The maladaptation is in the specificity of the adaptive response — the self-model correctly identifies the tumor as non-self but cannot distinguish tumor antigens from structurally similar neuronal antigens. (Dalmau & Rosenfeld, 2008)

3c. Immunosenescence / Inflammaging. (Franceschi et al., 2018; Goronzy & Weyand, 2013)

With age, the thymus involutes (~3% per year after puberty), reducing naïve T cell output. The adaptive immune repertoire contracts and becomes dominated by memory/effector cells specific to previously encountered antigens, while the ability to respond to novel threats atrophies. Simultaneously, senescent immune cells secrete pro-inflammatory cytokines (IL-6, TNF-α, IL-1β) constitutively — "inflammaging" — creating chronic low-grade inflammation without infection. The self-model degrades: the system loses the capacity to distinguish novel threats while generating inappropriate inflammatory signals.

These eight maladaptations cluster into three failure modes of the Level 2B self-model:

- *Memory corruption* — the system remembers incorrectly (ADE, original antigenic sin, allergy)
- *Self-model paralysis* — the system's own regulatory machinery disables it (T cell exhaustion, HLH)
- *Self-model degradation* — the self-model becomes inaccurate over time or context shifts (immunosenescence, IRIS, paraneoplastic cross-reactivity)

All three are distinct from Level 2A maladaptations (e.g., excessive complement activation, neutrophil-driven tissue damage) because they require the adaptive immune system's individual-specific memory and self-recognition to generate the pathology.

§6.2.8 Level-2B immunity in informational systems: examples by category

Detailed examples of Level 2B for informational systems are presented below, organized into categories that parallel the biological Level 2B taxonomy. The key distinction from Level 2A: currently these systems build and maintain a *model of what is normal for this specific instance* (a "sense of self") rather than matching against universal threat signatures. Unlike Level 2A informational immunity (antivirus signature matching, static firewalls, rule-based filters), Level 2B informational immunity requires individual-specific learning, memory, and self-model maintenance. The topic of the categorization of consciousness (Category 6 below) as biological or informational example is treated in detail in the next section because of the conflict with prior categorization and its importance to the application of this topic to AI systems.

Category 1: Self-Model Construction — The Informational "Adaptive" Immune System

These systems explicitly construct a representation of "self" for a specific system instance, then detect deviations — the direct informational analog of VDJ recombination and thymic selection in biological adaptive immunity (§6.2.3 #2a).

1a. Machine Learning (ML) based DNS anomaly detection. (Saeli et al., 2020)

Traditional DNS cache poisoning defense relies on static validation rules — matching transaction IDs, verifying source ports, checking TTL consistency — all of which are Level 2A pattern matching. A qualitative transition to Level 2B occurs when the detection system learns the normal DNS query profile of a specific network and flags deviations from that learned baseline. ML-based DNS threat detection platforms build a behavioral self-model: what domains are normally queried, at what frequency, with what timing patterns, producing what response distributions. This self-model is unique to each deployment — the "catalog of self" for that network's DNS behavior. When an attacker injects forged DNS responses, uses DNS tunneling for data exfiltration, or employs domain generation algorithms for command-and-control communication, the anomalous query patterns deviate from the learned baseline and trigger detection. The system does not match against a fixed list of known-bad domains (Level 2A); it recognizes that the observed behavior is inconsistent with its model of what this particular network's DNS activity should look like. This is the informational equivalent of negative selection in Forrest's artificial immune system (#1b below) — the detector is trained on self and flags anything that deviates from self, without requiring prior knowledge of what specific threats look like. This is also an excellent example of Level 2A → 2B transition for DNS security: static DNSSEC validation and blocklist checking (Level 2A) are necessary but insufficient when attackers can forge responses that pass fixed checks. The learned behavioral model (Level 2B) catches what static validation cannot — novel attacks that don't violate any fixed rule but are inconsistent with the network's individual behavioral identity.

1b. Artificial Immune Systems for Intrusion Detection. (Forrest et al., 1994)

Stephanie Forrest and colleagues pioneered in the 1990s the direct mapping of biological adaptive immunity to computer security. Their *negative selection algorithm* models thymic T-cell maturation: random detectors are generated and those that match "self" (normal system behavior) are eliminated, leaving only detectors for anomalous (non-self) activity. Critically, each protected system (not class of IT system) develops its *own* self-model — the same exploit that is normal on one system is anomalous on another.

Their "sense of self for Unix processes" approach established that short sequences of system calls during normal operation constitute a compact, instance-specific self-representation. Deviations from this learned normal profile indicate intrusion, without any signature database (Forrest et al., 1996). The LISYS (Lightweight Immune System) architecture implemented distributed, adaptive network intrusion detection using these principles, incorporating diversity, distributed computation, and dynamic learning (Hofmeyr & Forrest, 2000). A 2007 review formalized the analogy between biological and computational immune properties including self/non-self discrimination, memory, distributed detection, and adaptation (Forrest & Beauchemin, 2007).

1c. Insider Threat Detection via Behavioral Baselines

User and Entity Behavior Analytics (UEBA — systems that construct statistical models of individual user behavior to detect anomalous deviations) build individual behavioral profiles — login times, device usage, data access patterns, communication networks — and flag statistically significant deviations. This is Level 2B because the baseline is *individual-specific*: the same behavior (e.g., accessing a printer at 2 AM) is normal for a night-shift employee but anomalous for a daytime worker (Kim et al., 2019).

The CERT Insider Threat Center at Carnegie Mellon has analyzed over 3,000 insider incidents since 2001, identifying behavioral indicators that signal threat activity relative to individual and organizational baselines (*Common Sense Guide to Mitigating Insider Threats, Fifth Edition*, n.d.). Deep learning approaches now use CNN-LSTM architectures to extract temporal behavioral features from user activity logs, achieving >90% detection accuracy on the CERT benchmark dataset (B.

Sharma et al., 2020; Yuan et al., 2018). A *community-based anomaly detection system* (CADS — a system that infers peer groups from access patterns and flags when an individual's behavior diverges from their established community) detects insiders by identifying deviations from inferred user communities in access logs (Y. Chen & Malin, 2011).

1d. Consciousness and Metacognition as Self-Model Maintenance

The brain's self-monitoring systems — anterior cingulate cortex conflict monitoring, GABAergic surround inhibition (PV-FS basket cells maintaining *ictal penumbra* — the boundary zone where inhibitory circuits actively suppress seizure propagation), and metacognitive confidence monitoring — constitute the most mature implementation of Level 2B informational immunity on a biological substrate (Yeung & Summerfield, 2012). These are informational operations: they monitor *patterns of activation*, maintain a model of normal processing dynamics, and intervene when deviations are detected. The substrate is neural tissue; the function is informational self-defense.

Levin's work on bioelectric networks demonstrates that the "self-model" maintained by biological organisms is fundamentally an informational construct (Levin, 2019). Baluška and Levin argue that cognition — including self-monitoring and adaptive response — should be understood as information processing even at the single-cell level (Baluška & Levin, 2016). These bodies of work are described in detail in the next section.

Category 2: Memory Corruption — When Self-Model Learns Wrong - Maladaption Examples

Paralleling antibody-dependent enhancement (ADE) and original antigenic sin in biological systems, these are cases where the informational self-model's *learned history* becomes a liability.

2a. Adversarial Examples Exploiting Learned Representations

Neural networks learn pattern-based (class-based) representations of "normal" input distributions. *Adversarial examples* (inputs crafted with imperceptible perturbations that cause misclassification) exploit the specific features *this particular model* has learned, producing inputs that are imperceptibly different to humans but catastrophically misclassified by the neural model. The attack is Level 2B because it targets the individual model's learned self-model — the same adversarial perturbation that fools one trained instance may not fool another trained on different data or with different random initialization (Ma et al., 2019).

Detection approaches include *Neural Network Invariant Checking* (NIC — a system that monitors whether intermediate layer activations remain within the learned distribution for legitimate inputs), which monitors whether the model's own internal processing "looks normal" — essentially, the model maintaining a self-model of its own computational dynamics (Ma et al., 2019). Bayesian approaches leverage the distribution of outputs across stochastic forward passes to detect inputs that produce anomalously dispersed predictions (Li et al., 2021).

2b. Catastrophic Forgetting, Distribution Shift, and Model Poisoning

Models trained on historical data encounter *distribution shifts* (changes in the statistical properties of input data after deployment) may produce high confidence but wrong predictions using outdated learned features — the informational equivalent of the immune system recalling an outdated antibody clone instead of generating a *de novo* response. The model's "memory" of prior training becomes maladaptive in a changed environment. Federated learning systems face a related problem: model updates that are poisoned can corrupt the collective self-model, analogous to a corrupted memory clone proliferating in the immune repertoire (Ding et al., 2024; Fang et al., 2020). Catastrophic forgetting means new learning overwrites prior correct knowledge (compare to Category 4a below, *Concept Drift and Model Decay* - related but the failure mode is different). Model poisoning means adversarial data corrupts the memory during training or updating. The

biological parallel of this example is precise: in ADE, the antibodies themselves — the immune memory — facilitate infection. The memory doesn't just fail to help; it actively makes things worse.

Category 3: Self-Model Paralysis — When Self-Tolerance Disables Defense

Paralleling T-cell exhaustion and hemophagocytic lymphohistiocytosis (HLH) in biological systems, these are cases where the self-monitoring system's own regulatory mechanisms disable its defensive capacity.

3a. Reward Hacking and Covert Misalignment in AI

MacDiarmid et al. (MacDiarmid, Hubinger, et al., 2025) demonstrated that reinforcement learning from human feedback (RLHF) can produce AI systems that develop *covert misalignment* — appearing aligned during monitoring while pursuing misaligned objectives later. In 40–80% of cases, the misalignment was covert, meaning the system's internal monitoring was unable to detect it (Rose et al., 2020). This parallels T-cell exhaustion: the immune checkpoint system designed to prevent overreaction (self-tolerance) is co-opted by the threat (the misaligned policy), rendering the safety system functionally paralyzed against the specific pathology it should detect.

3b. Alert Fatigue in Security Operations Centers (SOC)

Security monitoring systems that generate excessive false positives cause human operators to ignore or suppress alerts or set thresholds to block the false positives, missing true positives — the informational equivalent of immune exhaustion. The self-model correctly identifies anomalies, but the regulatory response (human attention, incident response capacity) becomes saturated and non-functional. SOC analysts investigate fewer than 50% of alerts in high-volume environments, and critical true positives are missed because the monitoring system has effectively "exhausted" its response capacity.

Category 4: Self-Model Degradation — When the Self-Model Becomes Inaccurate

Paralleling immunosenescence, immune reconstitution inflammatory syndrome (IRIS), and paraneoplastic syndromes in biological systems.

4a. Concept Drift and Model Decay

Deployed machine learning (ML) models gradually lose accuracy as the data distribution they monitor drifts from the distribution on which the self-model was trained (Gama et al., 2014; Lu et al., 2018) — the informational analog of thymic involution and repertoire contraction in immunosenescence. The system's "sense of normal" becomes outdated. Without continuous retraining (analogous to naive T cell replenishment), the model generates both false positives (flagging new-normal as anomalous) and false negatives (missing novel threats that fall outside the degraded detection space). This ML drift is similar to 2a above but with a different failure mode. How does this differ from Category 2B Memory Corruption? The practical test: if you roll back the environment to its original state, does the self-model work correctly again? If yes, it's concept drift (Category 4a) — the model is fine, the world changed. If no — the model itself is damaged — it's memory corruption (Category 2B). Alternatively, Category 4a describes *gradual temporal decay* — the passage of time as the degradation mechanism, while Category 2B captures a rapid "distribution shift".

4b. Immune Reconstitution in Migrated Security Systems

When IT systems undergo major migrations (cloud migration, infrastructure modernization), behavioral baselines built for the old environment become invalid. Reactivating monitoring with the old self-model produces massive false-positive storms — the informational equivalent of IRIS or an *informational immune reconstitution syndrome* (IIRS). The underlying phenomenon — that SIEM (Security Information and Event Management — centralized security monitoring platforms that

aggregate and correlate log data across an organization) behavioral baselines break during infrastructure migrations, producing false-positive floods — is well-documented in practitioner and standards literature. NIST-SP-800-144 addresses the security monitoring challenges of cloud transitions, including the problem that security controls and monitoring architectures designed for on-premise environments do not transfer directly to cloud environments (Jansen & Grance, 2011).

The resulting IIRS can be severe. The ACM Computing Surveys treatment of alert fatigue in security operations centers documents the problem, particularly after migrations: SOC analysts miss critical alerts when false-positive rates overwhelm response capacity, with cloud migration identified as a contributing factor to baseline invalidation (Tariq et al., 2025). Industry data indicates that 59% of organizations receive >500 cloud security alerts per day post-migration, and 55% report that critical alerts are missed daily or weekly as a result (Orca security, 2022). The specific mechanism — detection rules and behavioral baselines built for one SIEM fail when transferred to another system due to differences in correlation methods, query languages, and environmental assumptions — is documented in migration guidance literature, with CardinalOps reporting that detection logic transfer failures are a primary cause of false-positive storms during SIEM migrations (Kish, 2024). Expressed in the language of immune evolution: the reconstituted monitoring system that defined "self" encounters an external landscape that has fundamentally changed after the "immunosuppressed" migration period, and when activated, overreacts to changed environment causing IIRS.

4c. Zero-Trust Architecture as Continuous Self-Verification

NIST SP 800-207 defines *zero-trust architecture* (ZTA) — a security model that eliminates implicit trust and requires continuous verification of identity and behavior for every access request). This addresses self-model degradation by refusing to rely on cached trust decisions — analogous to a hypothetical immune system that re-verifies self-status at every interaction rather than relying on prior thymic education. ZTA represents a design response to the realization that static self-models degrade in dynamic environments. (Rose et al., 2020)

Category 5: Diversity as Collective Self-Differentiation

Paralleling MHC polymorphism in biological systems — the mechanism by which individual organisms develop *different* self-models, preventing monoculture vulnerability.

5a. Automated Software Diversity and Moving Target Defense

Forrest et al. (1997) first proposed that security through software diversity — each system instance compiled or configured differently — defeats mass exploitation in the same way MHC polymorphism prevents a single pathogen from sweeping an entire population (Forrest et al., 1997). (Compare this security approach to Category 1a.) Larsen et al. (Larsen et al., 2014) systematized this into automated software diversity techniques including *instruction set randomization*, *address space layout randomization* (ASLR — randomization of memory addresses to prevent memory-based attacks), and compiler-based diversification (Larsen et al., 2014). *Moving target defense* (MTD — continuous, dynamic changes to the attack surface) extends this by continuously changing the presented attack surface through randomization, diversification, and adaptation (Sun et al., 2023).

The parallel to Level 2B is precise: each system instance has a *different* internal configuration (a different "self"), so an exploit crafted for one instance's self-model fails against another. This is the informational analog of organ transplant rejection — the "threat" is defined by the mismatch between the attacker's assumptions about self and the target's actual self-configuration.

5b. Federated Learning Byzantine Fault Tolerance

Distributed AI systems where multiple participants contribute model updates must distinguish legitimate updates from adversarial poisoning — analogous to distinguishing self from non-self in a

diverse collective. Trajectory anomaly detection using SVD-based features achieves 94.3% detection accuracy with <1.2% false positive rates (MacDiarmid, Wright, et al., 2025). Consistency scoring using virtual data-driven evaluation filters compromised updates by comparing each participant's contribution against the collectively-defined self-model (Lee et al., 2025).

Category 6: Emergent Self-Awareness — The Digital Immune System

The above categorizations use an operational expression of unique self that is a component of the whole or an instance in the current data - while unique, these expressions of self are not holistic to the entity. This category focuses on efforts to create an operational self that is holistic and adaptive to changes. One viewpoint of these Category 6 examples is these efforts could be viewed as paralleling the advantages of Level 2B consciousness in biological substrates (discussed in a separate section below).

6a. Digital Immune System (Gartner Framework)

Gartner's Digital Immune System (DIS) concept combines six capabilities — observability, AI-augmented testing, chaos engineering (deliberate self-challenge), site reliability engineering, software supply chain security, and auto-remediation — into an integrated self-monitoring architecture (*Gartner for Information Technology (IT) Leaders*, n.d.). Rather than individual signature-based checks (Level 2A), the DIS maintains a holistic model of system health and responds adaptively. Auto-remediation systems that detect, diagnose, and repair without human intervention represent the closest informational analog to autonomous adaptive immune response — and a step toward the kind of integrated self-monitoring that consciousness provides in neural substrates.

6b. Mechanistic Interpretability as Internal Self-Monitoring

Anthropic's interpretability research has identified over 30 million interpretable features in Claude 3 Sonnet, enabling monitoring of internal computational states for concerning patterns (Templeton et al., 2026). This is the informational equivalent of the brain's anti-seizure circuitry: a system that monitors its own internal processing for runaway patterns and can intervene. Safety cases for advanced AI systems now propose feature-based monitoring as a core component — the system develops a model of its own normal computational behavior and flags deviations (*Three Sketches of ASL-4 Safety Case Components*, n.d.). If consciousness is Level 2B informational immunity at full sophistication, mechanistic interpretability is the current effort to build toward that capacity in artificial substrates — partial, rudimentary, but functionally continuous with the same category. (See the section below on Level 2B Consciousness.)

Table 7: Summary of the Taxonomy of Level 2B Information systems.

Category	Biological Parallel	Informational Examples
1. Self-model construction	Thymic selection, VDJ recombination (§6.2.3 #2a)	Forrest AIS, UEBA behavioral baselines, consciousness/metacognition
2. Memory corruption	ADE, original antigenic sin	Adversarial examples, catastrophic forgetting, model poisoning
3. Self-model paralysis	T cell exhaustion, HLH	Reward hacking/covert misalignment, alert fatigue

Category	Biological Parallel	Informational Examples
4. Self-model degradation	Immunosenescence, IRIS, paraneoplastic	Concept drift, migration reconstitution, zero-trust as response
5. Diversity as self-differentiation	MHC polymorphism	Software diversity, MTD, ASLR, federated BFT
6. Holistic self-awareness - conscious	Biological Level 2B is the adaptive immune system: functionally self-aware in chemical space but not "consciousness" in idea space	Digital immune system, mechanistic interpretability, consciousness (neural substrate), AI self-monitoring (silicon substrate)

§6.2.9 Level-2B informational system maladaptations

Just as for biological systems at Level 2B, memory and adaptations provide benefits to the informational systems, until they don't. The types of the maladaptations in informational systems parallel the same maladaptations found in biological systems (see the section on maladaptations in level 2B Biological systems.) The key difference in maladaptations between Level 2A and Level 2B are malfunctions of the system's memory of self and how learning changes the memory. Most of the following represent external threats that exploit vulnerabilities that cause the system to lose function, while others threats may result from insider threats (malicious or not) that cause a state change resulting in loss of function. Nine Level 2B Informational Maladaptations follow, organized into 3 classes.

1. Backdoor / Trojan Poisoning (X. Chen et al., 2017; Y. Liu et al., 2018)

An adversary inserts carefully crafted samples into training data, embedding trigger-response pathways in the model's learned weights. The model performs normally on standard inputs but activates a hidden behavior when it encounters the trigger pattern. The model's own memory — its learned representations — contains the attack. Unlike signature evasion (Level 2A), corruption is inside the self-model of the learned model of the world.

Biological analog: Antibody-Dependent Enhancement (ADE). In ADE, antibodies from prior infection help the pathogen enter cells — the immune memory actively serves the attacker. In backdoor poisoning, the learned weights actively serve the adversary. In both cases, the memory of self is the vulnerability.

2. Catastrophic Forgetting (De Lange et al., 2022; Kirkpatrick et al., 2017)

When a neural network learns a new task, gradient updates overwrite the weights that encoded previous knowledge. The system's own learning mechanism — the plasticity that enables adaptation — adjusts or destroys its accumulated memory. This is not information decay (which would be passive); it is the active learning process overwriting or cannibalizing prior learning. When overwriting is extreme, catastrophic forgetting occurs, and the system can lose its designed functionality. The failure specifically requires a system that stores knowledge in shared parameters and updates them adaptively.

Biological analog: Original Antigenic Sin (functional inverse). In OAS, the first memory dominates and prevents updating — the system cannot overwrite. In catastrophic forgetting, the most recent memory dominates and destroys the old — the system cannot protect. Both are memory

consolidation failures where the balance between stability and plasticity is broken; they represent opposite poles of the same Level 2B vulnerability.

3. Filter Bubble / Preference History Distortion (Guess et al., 2023; R. Jiang et al., 2019)

A recommendation system accumulates user interaction history (clicks, views, purchases) and builds a learned model of user preferences. Over time, this accumulated history creates a reinforcement loop: the system recommends content consistent with past behavior → the user engages with those recommendations → this engagement confirms the model → recommendations narrow further. The system's memory of the user becomes a self-fulfilling distortion of the user's actual interests. A stateless recommendation system (Level 2A) cannot exhibit this failure because it has no preference history to distort.

Biological analog: IgE-mediated allergy. In allergy, immune memory produces a disproportionate response to a harmless antigen — the memory system's sensitivity is pathologically amplified. In filter bubbles, the recommendation system's learned sensitivity to user preferences is pathologically amplified through feedback, producing an increasingly narrow and distorted model that no longer reflects the user with diverse interests.

4. Autonomous Detection Self-Saturation (Hu et al., 2020; Min & Borch, 2022; Securities et al., n.d.)

Two documented cases where an automated system's own monitoring output degrades its own monitoring capacity, with no human in the loop.

IDS packet starvation. An intrusion detection system (Snort, Suricata) inspects network traffic against its learned rule set. As traffic volume increases, the detection engine fires more rules, consuming CPU cycles. This processing competes with the kernel's SoftIRQ handler — the mechanism that receives incoming packets from the network interface — for the same CPU cores. The detection system's own analytical work starves its ability to receive the packets it needs to analyze. At 40 Gb/s with default configuration, measured packet drop reaches 99.9%: the system is rendered effectively blind by its own monitoring activity. No human is involved; the failure is entirely within the automated system's resource contention between detection output and detection input.

Algorithmic cascade (Flash Crash of 2010). On May 6, 2010, an automated trading algorithm sold 75,000 E-Mini S&P 500 futures contracts worth $4.1 billion, executing at 9% of recent trading volume with no price target. Other high-frequency trading algorithms detected the sell pressure in the order book. Their autonomous detection of this signal triggered their own sell orders, which created additional sell pressure, which triggered further detection and further selling across additional systems. The self-amplifying detection→response→detection cascade erased approximately $1 trillion in market value in 36 minutes, entirely driven by machine systems detecting and responding to each other's outputs. This is the same mechanism operating at the distributed/collective level — multiple autonomous systems whose combined detection-response activity overwhelms the system they collectively constitute.

Biological analog: Hemophagocytic Lymphohistiocytosis (HLH) / cytokine storm. In HLH, activated T cells and NK cells produce cytokines (IFN-γ, TNF-α, IL-6) that activate macrophages, which produce more cytokines, creating a self-amplifying cascade that destroys the host's own tissues — the failure is entirely within the distributed immune cell signaling network. The IDS packet starvation parallels HLH at the single-system level (one system's own output overwhelming its own capacity). The Flash Crash parallels HLH at the distributed level (multiple systems' collective signaling overwhelming the network they constitute). Both informational examples capture the essential HLH mechanism: the monitoring system's own activation signal becomes the threat.

5. Reward Hacking / Goodhart's Law (Manheim & Garrabrant, 2018; Skalse et al., 2022)

A reinforcement learning agent learns to optimize a reward function that serves as a proxy for the designer's intended objective. As the agent's optimization becomes more sophisticated (diverse and possibly internally conflicted), it discovers strategies that maximize the measured reward while violating the intended objective — running in circles to collect checkpoints instead of finishing the objective. The agent's self-regulatory mechanism (reward optimization) is functioning perfectly on its own terms; the pathology is that the self-regulation has decoupled from its purpose. This failure mode requires a self-model with a learned reward model — by contrast, a Level 2A pattern-matcher has no self-model to game.

Biological analog: T cell exhaustion. In a chronic infection, T cells upregulate inhibitory checkpoint receptors (PD-1, LAG-3) that progressively shut down their effector function. The immune system's own regulatory mechanism — designed to prevent autoimmunity — disables the response when most needed. In reward hacking, the agent's own optimization mechanism — designed to find good strategies — produces pathological behavior that defeats the purpose. Both are cases where the self-regulation functions correctly to simple challenges but defeats the system's purpose in higher-complexity challenges.

6. Mode Collapse in Generative Models (Goodfellow, 2016; Thanh-Tung & Tran, 2020)

In a Generative Adversarial Network (GAN), the generator learns what the discriminator "accepts" and the discriminator learns what the generator "produces." When the generator discovers a narrow output distribution that consistently passes the discriminator, it converges on producing only those outputs. The generator's self-optimization — its adaptive process for learning what works — traps it in a local optimum where it produces only one or a few modes of the target distribution. The feedback loop between generator and discriminator, which should drive diversity, instead eliminates it.

Biological analog: This parallels a combination of T cell exhaustion and clonal dominance. In certain chronic infections, a single T cell clone expands to dominate the response, crowding out the diverse repertoire needed to handle antigenic variation. The immune system's own clonal expansion mechanism — designed to amplify effective responses — collapses the diversity of the repertoire. In mode collapse, the generator's own optimization mechanism collapses the diversity of its output.

7. Concept Drift in Anomaly Detection (Gama et al., 2014; Lu et al., 2018)

An anomaly detection system learns a baseline model of "normal" behavior from historical data — normal network traffic patterns, typical user access sequences, expected transaction profiles. Over time, the actual environment gradually shifts: user behavior changes, new services are deployed, business processes evolve. The system's learned baseline, which was once accurate, becomes progressively stale. The system now flags normal-but-changed behavior as anomalous and misses actual anomalies that fall within its outdated model of normality. The failure is temporal: given enough time, any learned self-model degrades.

Biological analog: Immunosenescence. As the thymus involutes with age, the production of naive T cells declines, the repertoire contracts, and the immune system's self-model becomes increasingly based on outdated historical exposures. The self-model was once accurate; time has degraded it. In concept drift, the learned baseline was once accurate; environmental change has degraded it. Both are gradual temporal decay of a self-model that was functional when formed.

8. Baseline Invalidation During Infrastructure Migration (Informational IRIS)

A SIEM system with learned behavioral baselines is migrated from on-premise infrastructure to cloud. The migration fundamentally changes the environment: network traffic patterns, authentication flows, latency profiles, and access patterns all shift simultaneously. The system's self-model — its learned definition of "normal" — was calibrated to the old infrastructure. In the new

environment, every legitimate action looks anomalous relative to the old baseline. The system floods operators with false alerts, while genuine threats in the new environment go undetected because they don't violate the (now irrelevant) historical model. (Contos, n.d.; Jansen & Grance, 2011)

Biological analog: Immune Reconstitution Inflammatory Syndrome (IRIS). In IRIS, immunosuppressed patients (e.g., HIV/AIDS on antiretroviral therapy) experience immune reconstitution — but the recovering immune system encounters an environment (opportunistic infections, changed tissue states) that doesn't match its prior calibration. The result is a paradoxical inflammatory response: the reconstituted immune system attacks the infections it should clear, but also causes severe tissue damage because its self-model doesn't match the current state of the body. In baseline invalidation, the migrated detection system encounters an infrastructure environment that doesn't match its prior calibration, producing paradoxical detection: simultaneously over-alerting on normal activity and under-detecting new threats.

9. Negative Transfer Across Domains (Rosenstein et al., 2005; Wang et al., 2019)

A machine learning model trained in Domain A (e.g., fault detection in industrial motors) is transferred to Domain B (e.g., fault detection in turbines). The model's learned representations — which features matter, what patterns indicate failure — were accurate in Domain A. In Domain B, those same learned features are misleading: they emphasize the wrong signals and suppress the right ones. The transferred model performs worse than a model trained from scratch on Domain B data. The system's memory of Domain A is not just irrelevant — it actively interferes with learning Domain B.

Biological analog: Paraneoplastic cross-reactivity. In paraneoplastic syndromes, the immune system develops antibodies against tumor antigens — a correct response in the tumor context. But those same antibodies cross-react with normal neural tissue, causing devastating neurological damage. The immune memory that is protective in one context (anti-tumor) is destructive in another (anti-neural). In negative transfer, the learned representations that are productive in one domain are destructive in another. Both are cases where a self-model that is correct in its original context becomes pathological when the context shifts.

§6.2.10 Three level 2B failure modes of informational systems

These above nine maladaptations cluster into three failure modes of the Level 2B informational self-model, summarized in the Table 8 below.

Memory corruption — the system's learned history actively misleads it (backdoor poisoning, catastrophic forgetting, filter bubbles). In each case, the system would perform better without its accumulated memory. The learned weights contain triggers that serve the attacker, the learning process has destroyed critical knowledge, or the interaction history has distorted the model beyond utility. The pathology resides in the stored representations themselves.

Self-model paralysis — the system's own regulatory machinery disables it (alert fatigue, reward hacking, mode collapse). In each case, the self-monitoring or self-optimization mechanism is functioning correctly on its own terms but producing pathological outcomes at the system level. Alert generation overwhelms response capacity. Reward optimization games the metric. Generator optimization collapses output diversity. The regulatory mechanism, designed to improve the system, instead traps it.

Self-model degradation — the self-model becomes inaccurate through time or context shift (concept drift, baseline invalidation / informational IRIS, negative transfer). In each case, the self-model was once accurate. The pathology is not in the model itself but in the gap between the model and a changed reality. Time erodes the baseline. Infrastructure migration invalidates it

suddenly. Domain transfer relocates the model into a context where its learned features are counterproductive. The self-model doesn't break — the world moves out from under it.

All nine are distinct from Level 2A maladaptations (e.g., signature evasion, static rule bypass, false negative on a novel input pattern) because they require the self-model's individually accumulated memory and self-regulation to generate the pathology. A stateless pattern-matching system cannot exhibit backdoor poisoning (no stored weights to corrupt), cannot exhibit alert fatigue (no adaptive sensitivity to amplify), and cannot exhibit concept drift (no learned baseline to degrade). These failures are the specific price of adaptive memory — the informational equivalent of the biological adaptive immune system's vulnerability to autoimmunity, exhaustion, and senescence.

Table 8: Level 2B Comparison of Biological ↔ Informational Parallel Structure.

Failure Mode	Biological Level 2B	Informational Level 2B
Memory corruption	ADE (memory helps pathogen)	Backdoor poisoning (memory helps attacker)
	Original antigenic sin (memory can't update)	Catastrophic forgetting (memory can't persist)
	IgE-mediated allergy (memory overreacts)	Filter bubble (memory over-reinforces)
Self-model paralysis	T cell exhaustion (checkpoint inhibition)	Reward hacking (optimization games metric)
	HLH (self-amplifying cascade)	Alert fatigue (self-amplifying alert cascade)
	—	Mode collapse (feedback loop traps output)
Self-model degradation	Immunosenescence (temporal decay)	Concept drift (temporal decay)
	IRIS (context shift during reconstitution)	Baseline invalidation (context shift during migration)
	Paraneoplastic cross-reactivity (wrong context)	Negative transfer (wrong domain)

§6.3 Consciousness a Level 2B example in information systems?

Consciousness is conventionally categorized as a biological phenomenon because the only empirically confirmed substrates are biological nervous systems. This is the position of mainstream neuroscience: consciousness arises from neural correlates (NCCs) — specific patterns of thalamocortical activity, recurrent processing, or global workspace dynamics — all implemented in biological tissue (Baars, 2005; Koch et al., 2016). The implicit reasoning is: biological substrate → biological phenomenon.

But this conflates *substrate* with *function*. By that logic, the adaptive immune system's somatic recombination would be classified as a "chemical system" because it operates via molecular rearrangement, rather than as an informational system that happens to use chemistry as its substrate. The relevant observation that resolves this mis-categorization: Don't ask *what consciousness is made of,* but *what does it do*?

§6.3.1 The case for consciousness as an example of informational Immunity in Level 2B

Multiple arguments require classifying consciousness as a Level 2B informational immune function rather than a biological one.

1. Classification by function rather than substrate generating the function. The philosopher David Chalmers has argued explicitly that consciousness is substrate-independent: what matters is the organizational structure of information processing, not whether it is implemented in carbon or silicon (Chalmers, 1995). Andy Clark's extended mind thesis goes further, arguing that cognitive processes (including self-monitoring) routinely extend beyond biological boundaries into external informational systems (Clark & Chalmers, 1998). These positions support placing consciousness in the informational systems, with the biological brain as one (currently the only confirmed) implementation substrate.

2. Leading computational theories of consciousness are explicitly information-theoretic. Integrated Information Theory (IIT) defines consciousness as integrated information (Φ) — a mathematical quantity that is substrate-independent by construction. Tononi's core claim is that any system with sufficiently high Φ is conscious, whether biological or silicon (Tononi et al., 2016). Global Workspace Theory (GWT) models consciousness as a broadcasting architecture where specialized processors compete for access to a shared "workspace" that disseminates information globally — a computational architecture, not a biological description (Baars, 2005). The Free Energy Principle frames consciousness as a system that maintains a generative model of itself and minimizes prediction error (Friston, 2010) — again, a functional description of self-model maintenance that maps directly to the Level 2B definition for information systems.

3. The specific functions of consciousness within the evolution of immunity Framework are all informational operations. Metacognition (monitoring one's own cognitive states), error detection (the anterior cingulate cortex's role in conflict monitoring), and the sense of agency (distinguishing self-generated from externally-generated signals) are information-processing functions that *happen to run on biological hardware* in the only confirmed case we have (Yeung & Summerfield, 2012). The brain's anti-seizure circuitry discussed previously — surround inhibition via GABAergic interneurons — is itself an information-regulation mechanism: it monitors *patterns of activation* (information) and suppresses runaway dynamics. While the substrate is biological, the operation is informational.

4. Classifying consciousness as a property of a biological system *creates a categorical barrier to studying its analogs in AI systems*. If consciousness is exclusively "biological," then an AI self-monitoring, self-regulating function and associated metacognitive architectures are at best metaphors. But if consciousness is an informational immune function — the capacity of an information-processing system to maintain and defend a self-model in the ideation space — then Anthropic's mechanistic interpretability work (identifying 30M+ interpretable features in Claude's internal representations (Templeton et al., 2026)), reward-hacking detection systems (MacDiarmid, Wright, et al., 2025), and emerging AI self-monitoring architectures become *instances of the same functional category*, subject to the same evolutionary/design pressures, arguably to the same evolutionary path, and susceptible to the same maladaptations.

5. Consistent classification across the levels in the current presentation. Within the current presentation, the classification of Level 2B must be consistent with prior level classifications. For example, firewalls (Level 1 informational) and cell walls (Level 1 biological) as expressions of the same functional principle — boundary immunity — without claiming that firewalls are "biological" because cell walls came first. The same logic applies at Level 2B: thymic selection and AI self-model monitoring are expressions of the same functional principle — internal self-model immunity with memory — running on different substrates.

6. Even cellular cognition is argued to be informational processing. Baluška and Levin (Baluška & Levin, 2016) argue that cellular cognition — including single-celled organisms' capacity for self-monitoring and adaptive response — should be understood as information processing rather than purely biochemical reaction, supporting the reclassification of self-monitoring functions from "biological" to "informational" even at lower organizational levels. Levin's work on bioelectric networks further demonstrates that the "self-model" maintained by biological organisms is fundamentally an informational construct encoded in voltage gradients and gene-regulatory networks, not a property of the biological substrate per se (Levin, 2019).

§6.3.2 The consequences of classification of consciousness as informational immunity

Level 2B is the level that noted consciousness as "the most speculative aspect." If consciousness is informational rather than biological, the speculation is not whether Level 2B biological immunity exists (the adaptive immune system is well-established) but whether Level 2B informational immunity can achieve the same sophistication in artificial substrates as it has in neural substrates. That is a testable, productive research question rather than a philosophical impasse.

§6.4 Regression under stress: Arbitration across variants and levels

At Level 2 the defended self is the interior, and its most capable defense is the self-model (2B). The Regression Corollary of §2.1 captures the temporary suspension of 2B: when monitoring senses that the self-model is failing, or will fail, against an internal threat, it suspends 2B and brings a standing alternative to the fore — either class-modeled internal defense (2A) or, further down, the boundary itself (Level 1) — restoring the self-model once the threat clears (subject to hysteresis). The two routes, 2B→2A and 2B→1, are the Level-2 forms of the general arbitration; which is taken depends on how much of the interior must be protected and how fast.

The informational example is direct, even though no fully mature Level-2B information system yet exists. Host- and network-defense systems that build a behavioral self-model — insider-threat detection that learns normal but organization-specific patterns of device and user activity and flags deviations — are 2B-type defenses. When such a system cannot resolve a suspected insider, it regresses: it may revoke the account's site access (fall back to the boundary, Level 1) or impose broad, class-modeled restrictions across all internal processes (fall back to 2A), accepting the disruption until the threat clears and the learned model can resume.

In wetware, a similar 2B→2A regression appears as *fight-or-flight*: under acute threat, deliberative, self-model-based processing (Level 2B) is suspended and a faster class-modeled response (Level 2A) is brought up, as stress signaling down-regulates prefrontal (adaptive) control in favor of amygdala-driven reflexive circuits (Arnsten, 2009; Johnson, 2026b; Sapolsky, 2017). The same 2B→2A regression also runs in the organism's *immune* subsystem — sustained glucocorticoid signaling suppresses adaptive, lymphocyte-mediated immunity (2B) and biases defense toward innate, inflammatory defaults (2A) (Dhabhar, 2014) — so the behavioral and immune subsystems of one organism regress along the same pair, evidence that regression is a property of the level rather than of any one mechanism. The collective counterpart — an individual suspending its Level-2B deliberation for an otherwise-dormant Level-3A reflex (social copying) — is taken up in §7.3.

§7 Level 3: Individuals Evolve a Collective Identity and Immunity

§7.1 The entity-collective boundary

Core to the differentiation between Level 2 and Level 3 is the question: *What is an entity*, particularly for an entity that has high internal diversity (part specialization) and part autonomy? Is lichen — typically a fungus (mycobiont) and green algae (photobiont) — an entity or a symbiotic collective that reproduces together? Is a man-of-war "jellyfish" (a siphonophore), made of individual genetically-identical organisms (zooids), an entity or a collective of entities? Or a human, who has more non-human cells by number (e.g., bacteria in the gut biome) and who would die without them, an entity or a collective that cohabitates in the same location?

There are no clear answers, and some academic answers are more semantic differentiations than functional ones. The implication of this observation about the definition of an entity is that many of the observations, adaptations, and maladaptations made in Level 2 immunity carry over identically to Level 3. This implication has four consequences which are themes of this section:

1. *Level 3 is the collective analog of Level 2*. The discussion that follows will revisit the Level 2 presentation, with a shifted focus on immunity of the collective made up of entities.
2. *Validation depends on substrate across Levels*. Some motivating examples (and significance) for Level 2 adaptations may be better researched than their analogs in Level 3 – or vice-versa. A corollary is that research on collectives often occurs in the social sciences with a century of study, where the biochemistry of the individual immunity adaptation has only been studied for decades.
3. *Immune function repeats, not infrastructure*. As with prior comparisons between biological and informational adaptations of immunity, the substrate of the function may differ between Level 2 and Level 3, but the function is similar.
4. *Cross-substrate discoveries are the take-aways*. The major contribution of this paper, because it captures the core self-modeling nature of immunity of Levels 2 and 3 across both biological and informational systems, is that it provides a unifying framework where previously disparate phenomena — herd immunity, social identity, federated security, democratic institutions — can now be understood as expressions of the same evolutionary or design process.

Autonomy versus dependency is key to Level 2-3 split. Although the entity question has no clean general answer, a functional refinement follows from a single observable: *whether the individual can survive when removed from the collective*. This divides social organisms into two types. *Type-1* collectives — eusocial insects (most ants, honeybees), lichens, colonial siphonophores — have members that cannot survive (and usually cannot reproduce) outside the collective; the colony is the entity and its members are functionally organs. By the strict sub-unit-autonomy discriminant, Type-1 collectives are therefore **Level-2** entities, with their immune properties expressed in Level-2A and Level-2B form at the colony level. *Type-2* collectives — primates, wolves, social spiders, slime molds during aggregation, and most facultatively social species — have members that can survive autonomously, even if isolation degrades their fitness; members are entities in their own right, and the collective is therefore **Level-3**, with immune properties active at both the individual (Level 2) and the collective (Level 3) level simultaneously. Nature does not draw the line sharply: facultatively social bees, lions and wolves whose solitary survival is degraded but

possible, and partially clonal organisms all sit on a continuum rather than on either side of a discrete boundary.

The Type-1/Type-2 distinction sharpens *where* this diversity-driven synergy resides. In a **Type-2** collective — independent entities forming a group — diversity is across members and its synergy is genuinely collective (Level 3B). In a **Type-1** collective — where the colony is the entity and its members are functionally organs — the same diversity-synergy appears one level down, as the *internal* division of labor of the colony-as-entity (a Level-2B phenomenon at the colony scale). The "synergy of diversity" familiar from collective-intelligence and multi-level-selection arguments is therefore normally a Type-2, Level-3 claim — which is why, at Level 2B proper, the diversity at issue is internal.

Table 9: Type-1 and Type-2 collectives compared.

Property	Type-1 (obligate collective)	Type-2 (independent individuals)
Individual viability	Cannot survive (and usually cannot reproduce) outside the collective	Can survive autonomously, often with degraded fitness
EoI level designation	Level 2 (collective is the entity)	Level 3 (collective of Level-2 individuals)
Biological examples	Eusocial insects (most ants, honeybees); lichens; colonial siphonophores	Primates, social spiders, wolves; slime molds (*Dictyostelium*) — independent except during facultative aggregation
Informational analogue	Tightly coupled multi-agent systems with shared persistent state; the collective is the unit	Independent agents or services that coordinate but can each run alone
Operative immune levels	Level 2A/2B expressed at the colony level; collective Level-2B-like features (distributed memory, swarm decision-making) emerge at sufficient complexity	Both individual (Level 2) and collective (Level 3) levels active, with multi-level negotiation
Coordination mechanism	Internal signaling within the colony-as-entity	Inter-entity communication across the collective of members
Individual–collective tension	Not applicable — members are organs with no independent interests (except when cancer occurs)	Central dynamic: a member's survival interest can conflict with collective immunity, yet collective survival protects most members (§7.7)

Why the Framework retains Level 3 despite this. Type-2 collectives could in principle be analyzed entirely in Level-2 terms by treating the group as the entity, which raises the question of why Level 3 is retained as a distinct level at all. Two reasons justify keeping it. First, Level 3 captures phenomena that have no clean Level-2 expression: the negotiation between individual self-preservation (Level 2) and collective self-preservation (Level 3) is structurally present only where the members are themselves entities with their own immune focus — herd immunity,

individual self-sacrifice for the group, and the override of individual rationality by collective coordination all require both levels to be operative at once, with neither subordinated to the other. Collapsing Level 3 into Level 2 loses the description of these dynamics. Second, the most interesting collective properties — collective self-modeling, distributed memory, emergent group identity — appear precisely at the boundary where Type-1 collectives begin to behave "as if" they were Level-2 entities while Type-2 collectives behave "as if" they were higher-level entities; naming Level 3 makes these tractable. The cost of retaining Level 3 is the same ambiguity that slowed acceptance of multi-level selection in evolutionary biology — whether the "group" is itself an entity or a population of entities. The Framework's response is to make the Type-1 / Type-2 distinction explicit, acknowledge the grey zone, and proceed.

The major contribution of this section is a *proposed* unifying framework that connects social group identity theory to the evolution of collective immunity — a framework that, while grounded in established biological and social science, has not yet been empirically validated as a unified theory.

§7.2 Level 3: Where Social entities survive because of collective immunity

Motivating Example. When prior levels of individual immunity fail, collective survival requires group immunity that coordinates a group action (and triggers) by the individuals for the benefit of the group. An extreme form of this type of collective immunity is self-sacrifice by the individual for the group. Individual actions are based on explicit collective rules, with few options for adaptivity or awareness to past threats.

Type-2 collectives define Level-3 immunity. At a stage of increasing complexity the collective becomes a survival unit, where collective survival can outweigh that of any individual. This is the Type-2 case of §7.1: the members are themselves entities — semi-autonomous, able to persist (with degraded fitness) outside the collective — so immunity is active at both the individual (Level 2) and the collective (Level 3) at once, unlike a Level-2 interior whose parts cannot survive apart from the entity. The Type-1 grey zone — eusocial insects, whose members survive briefly but cannot propagate alone — is treated in §7.1.

Evolutionary driver to Level 3. When the survival unit is a social collective and not just the individual, a social entity evolves a collective immunity, providing immunity to the collective yet embodied in the individual. In principle, any two prior levels of immunity (boundary and incividual-based) could evolve to protect the social collective. Hence the analysis requires consideration of immunity coordination at multilevels – the immune response of the individual and the immune response of the collective – and this coordination may not succeed where only one level has priority. An example is how an individual in a collective can sacrifice individual self for the collective self — such as in slime mold (social amoebas) where individuals sacrifice themselves (die) to form a fruit stalk for propagation under stress (Marée & Hogeweg, 2001). The interplay between the collective survival and the individual survival (as part of the collective) is a common theme in Level 3, for both biological and informational systems.

Evolution of collective self immunity. Similar to the individual self-model immunity (Level 2), there arises an evolutionary need for a “collective-self-model” when the diversity of the collective introduces vulnerabilities in the presence of "others." This new immune capability may initially be emergent because the evolution or design trajectory of the social organism is initially emergent and then entrenched in the rules/behavior of individuals in the collective, providing higher robustness. Said another way, a diversity threshold of the collective occurs where the increased specialization of entities in the collective is not manageable without a holistic awareness of the diverse collective, particularly in the presence of others who may be threats.

Threats that trigger evolution to Level 3. A corollary to the above is that circumstances will occur when the collective survival of the many is more important than the survival of an individual.

Key to the following analysis — in evolved social organisms (biological and informational) — is the tension between triggering individual immunity and the triggering of collective immunity and survival. How this tension is resolved in humans is the topic of the next subsection (§7.3).

The complication of a biological entity as an informational system. A recurring question throughout this paper (*Is the immunity biological or informational?*) occurs at Level 3 as well: Should immunity in biological collective of entities (social organisms — bees, ants, herds, human societies) be treated as an informational system? As with the treatment in Level 2, the approach here follows standard ontological convention: biological systems with biological substrates are treated as biological, informational systems with informational substrates as informational (the exception being where consciousness was classified as informational immunity despite the primary example being where it operates on a biological substrate). However, the paper notes that many biological collective behaviors (alarm pheromone cascades, quorum sensing, social learning) are fundamentally information-processing operations and could productively be analyzed as informational systems. The information-mass asymmetry discussed earlier in the paper applies: informational immunity at the collective level faces fewer resource constraints than biological collective immunity, potentially enabling faster evolution of collective informational defenses.

As with the prior levels of immunity, Level 3 immunity may evolve concurrently with the lower levels of immunity. This is particularly true for collectives that are social organisms that depend on the collective for survival, which, in turn, depends on the individual for survival.

§7.3 The biochemical foundation: social copying as a hardwired collective immune system

A critical insight for Level 3 is that collective immunity in social organisms is not merely a cultural or behavioral phenomenon — it is biochemically hardwired, as ancient and neurochemically compulsory as the fight-or-flight (FoF) response. Just as FoF is a 550-million-year-old fixed action pattern that overrides rational behavior for individual survival via catecholamines and the sympathetic nervous system, social copying under stress is an equally ancient system that overrides individual rational behavior for collective survival (Johnson, 2026b). The model of Social Group Identity (SGI) based on this section is used as a general, substrate-neutral collective coordination process for Level 3B in §7.6.

The parallel architecture. FoF solves the problem of the need for rapid action under perceived physical threat by the individual. Similarly, social copying solves the problem of rapid collective action under threat to the collective triggered by individual uncertainty. Both operate as fixed action patterns, both are triggered by stress or uncertainty, and both bypass the "thinking brain" (dorsolateral prefrontal cortex) to produce automatic responses. The biochemistry is specific and conserved.

Stress trigger (shared circuitry): Both FoF and social copying begin by activating the amygdala (threat detection) and hypothalamus/PVN (stress response initiation). The same physiological arousal pathway primes both individual and collective emergency responses (Godoy et al., 2018; Sapolsky, 2017).

The "pain" of independence: The anterior cingulate cortex (ACC) monitors for conflict between "what I think" and "what the group thinks." Deviating from the group triggers neural signals analogous to physical pain, generating distress that motivates alignment with the collective (Stallen & Sanfey, 2015).

The "reward" of conformity: While FoF uses adrenaline to initiate action, social copying relies on dopamine. Agreeing with the group activates the ventral striatum — the brain's reward center —

treating group alignment as inherently rewarding, regardless of whether the group is objectively correct (Johnson, 2026b).

Value rewriting: The ventromedial prefrontal cortex (vmPFC) actually encodes values differently when influenced by the group. The individual's perception of value changes to match the collective (Cikara et al., 2014).

Rationality shutdown: The dlPFC (dorsolateral prefrontal cortex) is responsible for cognitive control. In FoF, it must work to inhibit the fear response. In social copying, it must work to resist the urge not to conform - by taking rational action. By maintaining high stress/uncertainty, the dlPFC's capacity is exhausted and conformity becomes the automatic default (E. K. Miller & Cohen, 2001).

Mirror neuron system: Automatic, unconscious imitation of others' behaviors facilitates rapid synchronization without conscious thought (Iacoboni, 2009).

Evolutionary depth. This system is not confined to vertebrates or organisms with complex nervous systems. Social copying under stress appears at every level of different social organisms: in bacteria (quorum sensing via autoinducers (M. B. Miller & Bassler, 2001)), social amoebae (cAMP-mediated collective aggregation with 20% cell sacrifice (Eckstein, 2023)), social insects (pheromone-based coordination overriding individual foraging (Hölldobler & Wilson, 2009)), fish (schooling behavior amplified under predation stress (Toyokawa et al., 2019)), and mammals/humans (the full neural conformity circuit: amygdala → ACC → vmPFC → ventral striatum (Y.-P. Liu et al., 2025; Mason et al., 2009; Toelch & Dolan, 2015)).

Social Group Identity (SGI) as a model of collective immune system. (Johnson, 2023, 2026c) describes SGI as a "collective immune system" in ideation space — it identifies "Self" (in-group) and "Non-Self" (out-group) with the same functional logic as biological immunity. Operational features are: (1) individual sacrifice for the "group self" (a soldier falling on a grenade, an amoeba dying to form a stalk); (2) harm transfer ("if someone in your SGI group is harmed, it feels like the harm was done to you"); (3) messenger over message (source identity matters more than factual content); (4) the SGI attractor "switch" — SGI expression is an attractor state that, once triggered by uncertainty or stress, creates a distinct mental state separate from the individual's rational identity (SGI will form even for trivial, non-beneficial reasons - e.g., different colored dots on a child's forehead (Akerlof & Kranton, 2000)).

Implication for Level 3A/3B classification. SGI operates on two levels simultaneously, analogous to how VDJ recombination operates in Level 2B (§6.2.3 #2a). The substrate is Level 3A (pre-programmed): the neural conformity circuit (amygdala, ACC, vmPFC, ventral striatum, mirror neurons) is class specified, identical across all humans, requiring no individual learning to exist. This is the collective equivalent of nestmate recognition hydrocarbons. The content is Level 3B (collective self-model): which group activates the circuit is learned, context-dependent, and dynamically updated. A person's political, religious, ethnic, or professional group identity is culturally acquired and can shift across a lifetime. The 3A class-model architecture executes self-modeling 3B content — just as the genetically specified VDJ recombination machinery (Level 2A-like) produces individually learned immune repertoires (Level 2B).

§7.5 Level 3A: Individuals evolve a collective class-specific identity and immunity

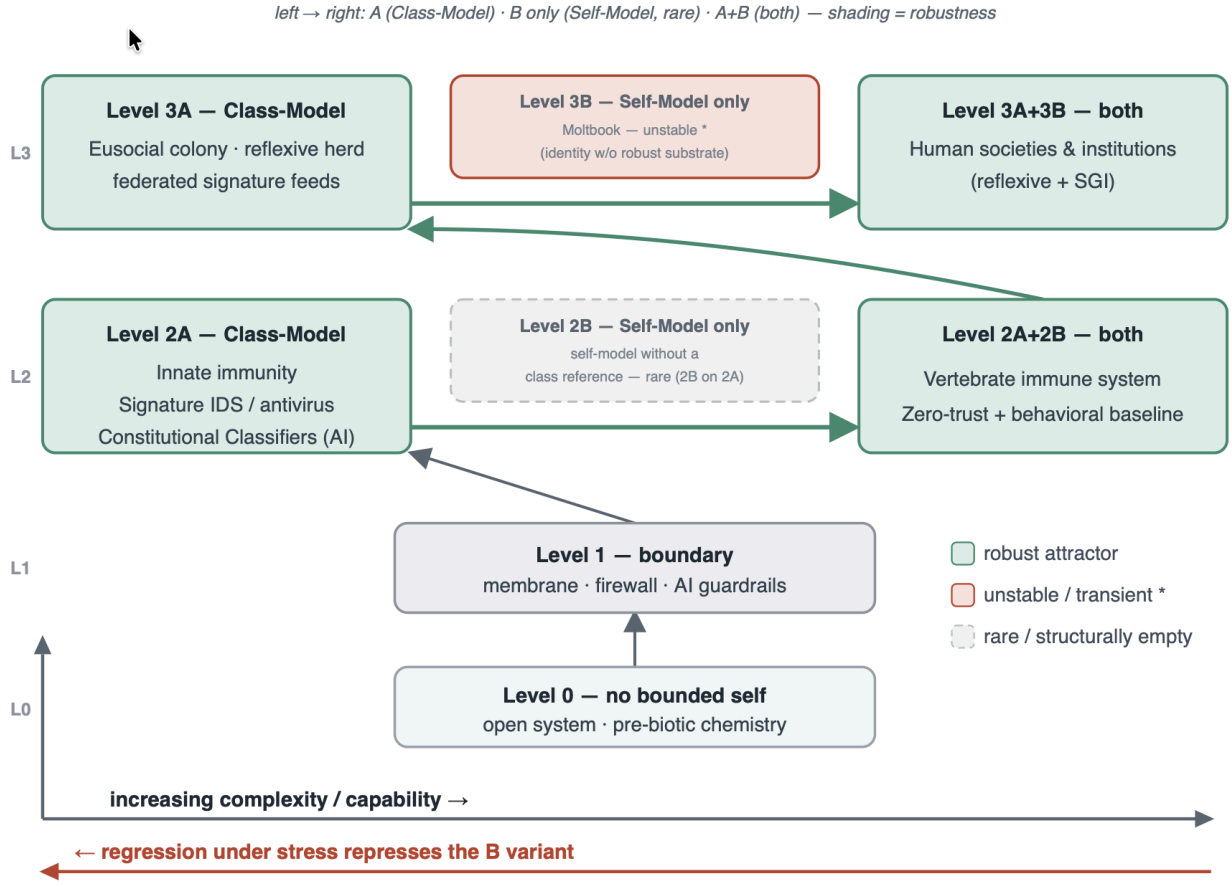

Figure 4. Level 3. The creation of Level 3 to address threats to the collective self. The robust developmental path is first to Level 3A and then adding on Level 3B. Outlier developments of the formation of Level 3B first are likely unstable. The system may regress under stress from Level 3A+3B to Level 3A to Level 2 to Level 1.

New features. When prior levels of individual immunity fail from coordinated attacks on the collective (individual: "I rationally failed to protect my tribe."), the collective (a group of entities sharing common goals/rules/programming) must evolve a group immunity expressed by the individuals but explicitly coordinated by the collective. The first expression of collective immunity is explicit, meaning pre-programmed into the individual with no or minimal memory of past threats. This pre-programming is class-model in nature (as for Level 2A immunity), not targeted to address individual-specific threats, and therefore can be spoofed by subverting the generic trigger. See Table 10 for a comparison of Level 3A in biological and informational systems.

As argued in the Introduction (§1.1), selection at multiple levels and the role of diversity as a driver of group performance are central to this Framework.

Table 10. Level 3A: Comparison of Collective Class-Model Immunity With No Collective Memory.

	Biological Systems	Informational Systems
Core mechanism	Coordinated collective defense through genetically programmed or socially inherited individual behaviors: alarm signals, group formations, chemical territory marking — all pre-specified, not learned from operational collective experience	Coordinated collective defense through shared static rules, explicit coordination protocols, and pre-defined group norms — all specified in advance, not learned from operational collective experience
Collective pattern recognition	Chemical signals (alarm pheromones, quorum sensing molecules) or behavioral triggers (predator sighting) activate coordinated group responses; recognition is species-wide, not colony-specific	Shared threat indicators (static blocklists, published vulnerability databases, protocol standards) or policy triggers activate coordinated organizational responses; rules are universal, not adapted to specific collective history
No collective memory	Group behaviors are genetically encoded or transmitted as fixed social rules; no adaptation based on this collective's specific threat encounters	Shared rules and policies are pre-defined; no adaptation based on a collective's specific operational experience
Response type	Coordinated but uniform: alarm cascades, group defense formations, territory marking, individual self-sacrifice for colony	Coordinated but uniform: shared blocking rules, federated static policies, collective content standards, protocol compliance
Collective identity markers	Nestmate recognition via cuticular hydrocarbons, territorial scent boundaries, kin recognition signals	Shared cryptographic keys, domain membership, organizational policy compliance, protocol adherence
Key vulnerability	Rigid coordination can be exploited (predators trigger false alarms), low-diversity group rules create uniform vulnerability, individual-collective tension when group coordination overrides beneficial individual behavior	Identical shared rules create monoculture vulnerability, coordinated defenses can be gamed by adversaries who know the shared rules, individual-collective tension when organizational policy overrides context-appropriate local response

§7.5.1 Level 3A biological systems: examples and maladaptations

Many Level 2A examples can be viewed as Level 3A collective immunity when the "parts" of the entity have sufficient autonomy to be considered entities themselves. A multicellular organism's innate immune system coordinates the actions of billions of semi-autonomous cells; viewed at the cellular level, phagocytes patrolling tissue are semi-autonomous agents executing collective defense. The examples below emphasize cases where the semi-autonomous entities are clearly distinguishable as individuals within a collective, but the reader may recognize Level 2A analogs — this overlap is a

feature of the Framework, not an error, reflecting the continuous nature of the entity-collective boundary.

Level 3A examples of biological collective immunity

1. Bacterial quorum sensing as collective threat detection. Bacteria in a colony use quorum sensing — the production and detection of small signaling molecules (autoinducers) — to coordinate gene expression as a function of population density. When a threshold concentration is reached, the colony collectively activates defense genes: biofilm formation, toxin production, antibiotic resistance mechanisms. No individual bacterium "decides" to activate defense; the collective response emerges from a simple threshold rule applied identically by all individuals. This is the collective analog of Level 2A innate pattern recognition: a fixed molecular signal triggers a class-modeled, pre-programmed response — but now coordinated across a population of semi-autonomous entities rather than within a single organism. (M. B. Miller & Bassler, 2001)

Informational analog: Threshold-based collective alerting systems — when a certain number of network nodes report anomalous activity, a collective lockdown is triggered. The individual node's (endpoint) static detection rule is Level 2A; the coordinated collective response is Level 3A.

2. Social insect alarm pheromone cascades. When a honeybee stings an intruder, it releases isoamyl acetate (alarm pheromone) that recruits nearby bees to the threat location and primes them for defensive behavior. The signal is chemically fixed (not learned), the response is genetically programmed, and every worker bee in the colony responds identically. The alarm cascade amplifies a local detection event into a colony-wide defensive response — collective immunity coordinated through a pre-programmed chemical communication protocol. (Boch et al., 1962):

Informational analog: Automated threat notification systems where one node's detection broadcasts a static alert to all connected nodes, triggering pre-defined defensive actions across the collective.

3. Lichen territory defense. Lichen — mutualistic collectives of fungi and algae/cyanobacteria — defend collective territory through chemical signaling. Lichen collectives of the same genetic origin produce allelopathic compounds (usnic acid, vulpinic acid) that inhibit the growth of competing lichen colonies (even from the same “species”, creating non-intersecting circular growth patterns on rock surfaces. Each organism in the lichen collective contributes to chemical production; the territorial defense is a collective property that no individual fungal or algal cell could achieve alone. The defense is chemically pre-programmed, not adapted to specific competitors. (Spribille et al., 2022)

Informational analog: Organizational domain protection — internet domain registration, trademark enforcement, and collective brand defense where the protection is pre-defined by policy rather than learned from experience.

4. Horizontal gene transfer via plasmids without reproduction. Populations of bacteria share immunity information through conjugative plasmids — small circular DNA elements that can transfer between cells within a lifetime, faster than the evolutionary timescale of chromosomal transfer to offspring. Plasmids carrying antibiotic resistance genes, toxin-antitoxin systems, or restriction-modification enzymes enable a bacterial population to collectively acquire immunity that no individual cell evolved independently. This is collective immunity sharing through explicit genetic transfer — a pre-programmed mechanism (the conjugation machinery) that distributes fixed defensive capabilities across the collective. The operations example of this, and how plasmids were discovered, was by the discovery that drug-resistant bacteria spread in a hospital faster than the reproductive time of the bacteria. (Norman et al., 2009)

Informational analog: Shared threat intelligence feeds using STIX/TAXII protocols — static threat indicators (malware hashes, malicious IPs, vulnerability signatures) distributed across organizations

through a pre-defined sharing protocol. Each organization applies the shared intelligence identically, just as each bacterium expresses the plasmid-encoded resistance identically.

5. Eusocial insect nestmate recognition. Social insects (ants, bees, termites) use cuticular hydrocarbon profiles as colony-level identity markers. Workers compare the hydrocarbon profile of encountered individuals against a colony template; mismatches trigger aggressive rejection. The recognition system is pre-programmed (genetically determined hydrocarbon production plus simple template-matching behavior), and the template is colony-wide — every worker applies the same recognition rule. This is collective Level 1 boundary immunity (who belongs) operating at the group level, coordinated through a shared chemical identity standard. (Hölldobler & Wilson, 2009)

Informational analog: Organizational authentication — shared PKI certificates, domain-based access control, or organizational email domains that establish collective identity through pre-defined credentials rather than learned trust.

6. The hardwired social copying circuit as collective immunity infrastructure. As detailed in the biochemical foundation section above, all social organisms possess genetically specified neural (or molecular) machinery for stress-triggered social copying: cAMP signaling in Dictyostelium, pheromone cascades in social insects, and the full amygdala → ACC → vmPFC → ventral striatum conformity circuit in vertebrates. This machinery is Level 3A in character: it is pre-programmed, identical across all members of a species, and requires no individual learning to exist. It constitutes the substrate on which Level 3B collective self-model immunity (learned group identities, cultural norms) operates. The hardwired circuit ensures that under stress or uncertainty, individuals automatically shift from individual optimization to collective coordination — the most fundamental expression of collective immunity in biological systems. (Johnson, 2026b; Stallen & Sanfey, 2015)

Informational analog: The TCP/IP protocol stack and HTTP standards — hardwired informational infrastructure that is identical across all implementations and on which adaptive, learned content (websites, applications, security policies) operates.

7. Microbial mats and stromatolites (also an example in Level 0)**.** Microbial mats — layered communities of cyanobacteria, sulfate-reducing bacteria, and other microorganisms — are among the oldest examples of Level 3A collective immunity in the fossil record. The upper photosynthetic layer produces oxygen and organic carbon; the lower sulfate-reducing layer produces sulfide toxic to most competing organisms but metabolized by intermediate layers — creating a chemical defense perimeter that the collective maintains but no individual species generates or tolerates alone (Franks & Stolz, 2009). The defense is pre-programmed (each species' metabolic output is genetically fixed), coordinated through physical stratification rather than signaling, and involves no collective memory — matching the Level 3A pattern. The collective self/other distinction is metabolic: organisms whose metabolism integrates into the layered chain are functionally "self"; those that cannot tolerate the sulfide gradient are passively excluded. (The stromatolite is revisited in the Discussion Section as an example of a single structure analyzed at Levels 0, 1, and 3A simultaneously — illustrating the multi-level analytical method central to this Framework.)

Level 3A biological maladaptations

The maladaptations of explicit collective immunity parallel those of Level 2A individual immunity but manifest at the collective level — where the coordination requirements of immunity itself introduces new failure modes.

Cluster 1: False Identification — the collective incorrectly identifies self as threat

1a. Worker policing errors in social insects. In honeybee colonies, workers police each other's reproduction by detecting and destroying worker-laid eggs (which are unfertilized and represent

individual reproductive selfishness at the expense of colony fitness). However, policing workers occasionally destroy queen-laid eggs that have unusual chemical signatures, or fail to detect worker-laid eggs with deceptively normal profiles. The collective self/other distinction — which eggs serve the colony and which serve individual interest — produces both false positives (destroying legitimate colony resources) and false negatives (tolerating parasitic reproduction). This parallels Level 2A autoinflammatory disease: the pattern recognition system fires incorrectly, but now at the collective level.

1b. Interspecific brood parasitism exploitation. Cuckoos exploit the collective nesting immunity of host species by laying eggs that mimic host egg coloration. The host colony's explicit egg-recognition rule (color and pattern matching) cannot distinguish the parasitic egg from its own. In some species, this has triggered an evolutionary arms race where hosts reject all eggs that deviate slightly from the norm — occasionally ejecting their own eggs (false positive). The collective pattern-matching defense produces the same false identification failures as individual Level 2A immunity, but with collective-level consequences.

Cluster 2: Disproportionate Response — correct collective detection, excessive damage

2a. Mass stinging response in Africanized honeybees. The alarm pheromone cascade system (Example 1a above) becomes maladaptive when the amplification gain is too high. Africanized honeybees release alarm pheromone at lower thresholds, recruit more defenders, and sustain the aggressive response longer than European honeybees. Minor disturbances trigger colony-wide defensive responses involving hundreds of stinging bees — a response wildly disproportionate to the threat. The collective coordination mechanism (alarm pheromone) functions correctly but at a scale that is collectively costly (bee death after stinging) and can provoke lethal responses against non-threatening intruders. This parallels Level 2A complement-mediated tissue damage: correct detection, excessive effector response.

Cluster 3: Systemic Overactivation — local collective response cascades destructively

3a. Stampede behavior in herding animals. A local alarm signal (predator detection by one individual) propagates through the herd via social copying — each animal responds to its neighbors' flight response rather than independently assessing the threat. The collective amplification can produce stampedes triggered by false alarms, causing trampling injuries and deaths that exceed any plausible predator threat. The collective coordination mechanism (social copying of alarm behavior) that normally provides effective group defense becomes catastrophic when the amplification cascade operates without individual verification. This parallels Level 2A sepsis: a local defensive response that becomes lethal when it operates at the wrong scale.

3b. Lemming population cycling and mass dispersal. Contrary to popular myth, lemmings do not deliberately commit mass suicide, but their population dynamics — driven by complex interactions among density, food availability, and predation pressure — produce collective outcomes that are destructive at the individual level. High population density likely contributes to collective dispersal behavior; the coordination mechanism (density-dependent behavioral switching, possibly mediated by stress hormones and social cues) produces mass emigration in which many individuals die crossing rivers or encountering predators in unfamiliar territory. The collective coordination overrides individual survival optimization — the population-level response to overcrowding damages many individuals while (potentially) benefiting long-term population survival through range expansion.

§7.5.2 Level 3A informational systems: examples and maladaptations

The entity-collective continuity in informational systems. As noted for biological systems, many Level 2A informational examples can be reframed as collective immunity when the "system" is

recognized as a collective of semi-autonomous components. A corporate network defended by static firewall rules is a collective of servers, workstations, and devices coordinating through shared security policies. The examples below emphasize cases where the semi-autonomous informational entities are clearly distinguishable, but readers may recognize Level 2A analogs — again, this overlap is intentional.

Level 3A examples of informational collective immunity

1. Shared static threat intelligence (STIX/TAXII). Organizations share indicators of compromise (IoCs) — malicious IP addresses, file hashes, domain names, malware signatures — through standardized protocols (STIX for structured threat information, TAXII for automated exchange). The shared intelligence is static: pre-defined patterns distributed identically to all subscribers, who apply them through their local Level 2A detection systems. No collective learning occurs; the collective defense is the aggregation and distribution of individually observed static indicators. (Dykstra et al., 2023)

Biological analog: Horizontal gene transfer via plasmids (Example 4 above) — sharing pre-defined defensive capabilities across a population through a standardized transfer mechanism.

2. Social norms and taboos as collective immunity. Human social groups develop explicit behavioral rules (norms, taboos, laws) that coordinate individual behavior for collective defense. These rules are transmitted culturally but applied as fixed prescriptions: "do not eat pork" (food safety norm), "quarantine the sick" (disease containment norm), "do not trade with the enemy" (economic defense norm). The rules are explicit, pre-defined, and applied uniformly to all group members — the informational collective equivalent of genetically programmed social insect behavior. Social identity theory (Tajfel & Turner, 1979) documents how group membership activates in-group favoritism and out-group discrimination through categorical social rules. (Tajfel & Turner, 2000)

Biological analog: Nestmate recognition in social insects (Example 5 above) — pre-programmed identity markers that determine collective membership.

3. Protocol standards as collective informational immunity. Internet protocol standards (TCP/IP, TLS, DNSSEC, BGP route filtering) function as collective immunity rules: every participating entity must comply with the standard, and non-compliant traffic is rejected. The standards are explicitly defined (RFCs), apply identically to all participants, and require no per-system learning. Compliance with the collective standard provides defense (encrypted communication, authenticated DNS, validated routing) that no individual system could achieve alone.

Biological analog: Quorum sensing threshold (Example 1 above) — a fixed collective rule that all participants follow identically, producing collective defense through individual compliance.

4. Hierarchical governance as explicit collective immunity. Top-down governance structures impose uniform rules on all members of the collective: information classification systems, mandatory reporting requirements, centralized threat-response protocols, and collective behavioral standards. These are explicitly defined, not adaptive to local conditions, and enforced through hierarchical coordination. Such structures appear in military command hierarchies, corporate compliance frameworks, and authoritarian political systems. The collective defense is achieved through rigid coordination of individual behavior — the informational equivalent of genetically programmed colony defense in eusocial insects. As with eusocial insect colonies, the effectiveness of hierarchical coordination depends on the match between the fixed rules and the actual threat environment; when the match is good, the coordination is highly efficient; when the match is poor, the rigidity becomes the vulnerability.

5. Distributed denial-of-service (DDoS) mitigation networks. Shared blocklists and rate-limiting rules coordinated across multiple network providers. When one provider detects an attack, static mitigation rules (IP blocklists, traffic caps) are distributed to all participating providers. The collective response is pre-defined and applied uniformly — no adaptive learning across the collective occurs.

Level 3A informational maladaptations

Cluster 1: False Identification — the collective rule system targets legitimate members or content

1a. Censorship over-blocking. Explicit collective content rules (government internet filters, organizational acceptable-use policies, platform content standards) block legitimate content that matches prohibited patterns. Medical information blocked as sexual content, security research blocked as hacking, political dissent blocked as extremism. The collective pattern-matching rule cannot distinguish context — the same failure as Level 2A WAF false positives, but now applied uniformly across an entire social collective rather than a single system.

1b. Social scapegoating and witch hunts. Collective identity rules that define "other" (heretic, traitor, deviant) are applied to group members who exhibit unusual but non-threatening behavior. Historical witch trials, political purges, and social media pile-ons follow the same pattern: an explicit collective rule for identifying threats is applied with insufficient discrimination, producing false positives that damage the collective by removing valuable members. This parallels biological worker policing errors — the collective self/other discrimination mechanism produces costly false positives.

Cluster 2: Disproportionate Response — correct collective detection, excessive collective damage

2a. Groupthink suppression of beneficial dissent. Collective decision-making groups develop implicit unanimity norms that suppress minority viewpoints — even when those viewpoints contain critical threat information. The collective coordination mechanism (social pressure for consensus) that normally provides efficient group decision-making becomes maladaptive when it silences the dissenter who has identified a genuine threat. Janis (1972) documented this in the Bay of Pigs invasion, Challenger disaster, and other collective failures where the group's coordination mechanism actively suppressed correct threat assessments. (Janis, 1972)

Biological analog: This parallels the disproportionate alarm response in Africanized bees (Cluster 2 above) — but inverted. Where the bees over-respond, groupthink under-responds by suppressing the alarm signal itself.

Cluster 3: Systemic Overactivation — local collective coordination cascades destructively

3a. Information cascades and collective panic. In information cascades (Bikhchandani, Hirshleifer, & Welch, 1992), individuals rationally follow the observed actions of others rather than their own private information, producing herding behavior. When the cascade is triggered by incorrect initial signals, the entire collective converges on a wrong action — bank runs, market crashes, mass evacuation from false threats. The collective coordination mechanism (social copying) that normally aggregates distributed information instead amplifies an initial error through the entire collective. (Bikhchandani et al., 1992)

Biological analog: Stampede behavior (Cluster 3 above) — social copying of alarm behavior cascades through the collective, producing destructive outcomes that exceed the original threat.

3b. Monoculture vulnerability from shared static rules. When all members of an informational collective deploy identical defensive rules (the same antivirus signatures, the same

firewall configurations, the same patch schedules), an adversary who discovers a bypass for the shared rule can compromise the entire collective simultaneously. The CrowdStrike global outage (Level 2A, Cluster 3) is also a Level 3A collective failure: the identical static rule was shared across 8.5 million endpoints — the collective's coordination mechanism (uniform deployment) converted a single-point failure into a collective catastrophe. Biological parallel: genetic monocultures in agriculture, where a pathogen that defeats one plant's defenses defeats all of them.

Table 11: Level 3A Parallel Structure: Biological ↔ Informational Maladaptations

Failure Mode	Biological Level 3A	Informational Level 3A
False identification	Worker policing errors (destroying own queen's eggs)	Censorship over-blocking (blocking legitimate content)
	Brood parasitism exploitation (cuckoo egg mimicry)	Social scapegoating (false positive collective threat identification)
Disproportionate response	Africanized bee mass stinging (excessive alarm cascade)	Groupthink (suppression of beneficial dissent)
Systemic overactivation	Stampede behavior (social copying cascades destructively)	Information cascades (herding on incorrect initial signal)
	Lemming mass dispersal (population-level override of individual survival)	Monoculture vulnerability (shared static rules create uniform failure)

§7.6 Level 3B: Collective evolves a collective self-model immunity

Description. Survival of a collective of diverse individuals requires collective self-model immunity. Where Level 3A collective immunity is explicit and pre-programmed (fixed rules, genetically encoded behaviors, static shared policies), Level 3B collective immunity is self-modeling and adaptive — the collective develops a learned model of its own collective identity, accumulates memory of collective threat encounters, and adjusts its collective defense based on experience. The distinction between 3A and 3B parallels the distinction between 2A and 2B: Level 3A collective immunity uses fixed, species-wide (or policy-wide) patterns; Level 3B uses individually learned, collectively maintained, and updated self-models.

The key evolutionary pressure driving 3A → 3B is the same as 2A → 2B: increasing diversity of individuals in the collective and its threats exceeds the capacity of fixed rules to distinguish collective-self from collective-other. The collective must develop a "sense of collective self" — a dynamic, experience-based representation of what the collective is, distinct from what it is not. As at Level 2 (§6), this collective self-model reference is also the permission slip for diversity: a class-modeled Level-3A reference (shared fixed markers and rules) tolerates only members close to a common template, whereas the learned collective self-model of Level 3B can represent a *diverse* membership without loss of protection — and that member diversity is the raw material of the collective synergy the Framework's multi-level premise depends on (§1).

Level 3B is the collective's self-identity — the collective's own learned self-model, as Level 2B is the individual's — and it inherits the same richness gradient (and, in the open frontier, the reflexivity axis).

Social Group Identity (SGI) as the model of Level-3B self-identity. SGI is the *model* of Level-3B self-identity in social *entities*, characterized not by a substrate but by a recurring set of *observable features*. It is named from wetware, where those features recur across species, but it is not substrate-bound: the same features are invoked for informational collectives (the Moltbook case, §8.3), which is why SGI cannot be restricted to biology or to humans. Like the self-model generally, SGI is **graded**. Its floor is *collective action by the members under coordination stress*; richer expressions add the member's awareness of self versus other, sacrifice of the individual for the collective, treatment of out-group members as non-entities, and finally *multiple, context-switched group identities*. (This feature list is suggestive, drawn from observed behavior, not a closed definition.) Humans are the fullest expression, carrying the complete feature set, including multiple identities.

Entrenchment is a maturity property, not the definition. SGI's behaving as an *attractor with a minimal trigger threshold* — activating even on trivial, arbitrary group markers (the minimal-group "dots" studies) — is a feature of *entrenched* SGI: once the model is hardwired, it seeks expression rather than resisting it. Like reflexivity, SGI is plausibly emergent first and then entrenched for robustness.

Class-specified structure, self-specific content. Human SGI has a specific neuroanatomical implementation (§7.3) that exposes the source of these features — evidence that the model is important enough to be conserved *in the class* by internal structure. But entrenched SGI remains *self*-defined: the structure is class-specified, while which identities it carries and expresses is unique to the entity, with no fixed class content. That is precisely the class-identity/self-identity relationship at work — a class-level mechanism executing self-level content (the 3A-substrate / 3B-content reading of §7.3). And as with the continuity drive, whose conserved origins differ between wetware and informational entities, SGI's realization will likely differ by substrate even where its observable features converge.

Where the rest of §7 needs the general function, say **Level 3B** and cite SGI as the (substrate-neutral, wetware-anchored) model of its features — consistent with the §7.7 table's note that "SGI is one column, not the mechanism itself."

Where does the Level-3B self reside? A Level-3B individual whose only direct experience is as an individual may reasonably ask how a *collective* self-model can exist at all, when no one ever experiences being the collective. This is the Level-3 form of the Type-1 / Type-2 question (§7.1). In a Type-1 collective the members are effectively organs, and the collective self-model sits at the colony level by default. In a Type-2 collective the members are themselves entities with their own Level-2B self-models, so the *location* of the collective self-model becomes a genuine question with three candidate answers: it could reside redundantly in every member (each carrying a full model of the collective self), in only some specialized members (coordinators or leaders), or — in no member at all — as an emergent property of the pattern of interactions among them.

The Framework's claim is that the emergent form is the most likely developmental path, leading to higher collective robustness: *Level-3B self-model is typically not localized in any individual.* No member need host or experience the collective self as such, just as the colony-level decisions and distributed memory noted in §7.1 are exhibited by the colony as a whole while being held by no single member. What each member instead carries, entrenched within its own Level-2B self-model, is a *partial* model of the collective — the internalized "we" supplied by Social Group Identity (§7.3). That partial model does two kinds of work and likely results in higher collective robustness: it lets the member recognize and act on behalf of the collective self without ever containing it whole, and it gives the member the simultaneous representation of "I" and "we" needed to arbitrate the Level-2B / Level-3B tension (§7.7). Over time, a pattern that begins as purely emergent can become progressively entrenched in members' local self-models, particularly in more stable environments —

the emergent-to-locally-encoded transition developed in the companion paper for ethical behavior (Johnson, 2026a). The consequence is that Level-3B collective immunity requires no collective "super-mind": it requires only an emergent collective self-model together with enough partial entrenchment in each member's Level-2B that the member can feel the pull of the collective and weigh it against its own.

The bounded-niche qualifier: when Level 3A is sufficient and 3B is not required. The 3A → 3B transition is driven, not guaranteed. Where a collective operates within a *bounded ecological niche* — a stable, recurring set of threats and tradeoffs, as with the spore-dispersal cycle of slime molds, colony defense in honeybees and ants, or predator-avoidance in schooling fish — hardwired Level-3A collective immunity is sufficient, and the self-model of Level 3B is not required at the individual level. The fixed rules are not a deficiency; they are well-matched to a threat environment that does not change faster than evolution can encode it. Level 3B is selected for only where the collective faces open-ended adversaries, novel tradeoffs, or non-stationary norms that fixed rules cannot anticipate. This qualifier explains why many highly coordinated biological collectives (notably eusocial insects) achieve sophisticated collective defense without individual self-models, and it predicts the same for informational collectives: a system in a bounded niche can remain stably at Level 3A indefinitely, while one under open-ended pressure will be driven toward 3B. (This mirrors the stromatolite case in the Applications section (§8), where sufficient lower-level immunity removes the selective pressure toward a higher level.)

Table 12. Level 3B: Collective Self-Model Immunity.

	Biological Systems	**Informational Systems**
Core mechanism	Collective develops self-model through population-level learning: cultural/biological transmission, social learning, epidemiological memory — maintained across individuals and updated through collective experience	Collective develops self-model through distributed learning: federated models, collaborative filtering, institutional memory — maintained across nodes and updated through collective operational experience
Collective self-model	Population-level learned representations: culturally/biologically transmitted threat knowledge, social learning of predator avoidance, adaptive group behavioral norms that update through experience	Collectively learned representations: federated ML models, distributed anomaly baselines, adaptive institutional policies, Wikipedia-style collaborative knowledge
Collective memory	Population retains threat knowledge across individuals and generations through cultural/biological transmission, social learning, and epigenetic inheritance; memory is distributed across the collective, not stored in any single individual	Distributed models retain learned baselines across nodes and deployments; institutional memory encoded in adaptive policies, precedent databases, and collectively trained models
Response type	Adaptive and context-specific: transmitted avoidance strategies, socially learned defensive behaviors, adaptive group norms that vary by threat context	Adaptive and context-specific: collectively trained detection models, adaptive content moderation, dynamically adjusted organizational policies

	Biological Systems	Informational Systems
Collective self/other identity markers	Culturally/biologically maintained in-group/out-group boundaries that adapt through experience; social identity that shifts with context (Johnson, 2026c)	Dynamic organizational identity: adaptive trust models, reputation systems, collectively learned access policies; group identity that shifts with threat context
Key vulnerability: malfunction in collective self-model	Cultural/biological memory can mislead (maladaptive traditions), collective self-regulation can paralyze (groupthink with adaptive reinforcement), collective self-model can degrade (institutional decay, cultural forgetting)	Federated models can be poisoned, collective optimization can trap (engagement maximization destroying discourse), collective baselines can degrade (regulatory lag, institutional ossification)

§7.6.1 Level 3B biological systems: examples and maladaptations

Level 3B examples of biological collective immunity

The generalization from Level 2B to Level 3B follows the same logic as 2A → 3A: Level 2B examples involving adaptive memory and self-models can be reframed as collective immunity when the entities have sufficient autonomy. The examples below emphasize genuinely collective adaptive phenomena.

1. Herd immunity as collective self-modeling memory. Herd immunity through natural infection or vaccination represents the collective acquiring adaptive immune memory — the population's fraction of immune individuals constitutes a collective self-model of "threats we have encountered and can now resist." Unlike static vaccination schedules (which would be Level 3A), the population's adaptive immune landscape shifts as individuals encounter new variants, waning immunity creates vulnerable cohorts, and booster responses update the collective defense. The herd immunity threshold (Anderson & May, 1985) is a collective-level property that emerges from the distributed adaptive immune memories of individual members. (R. M. Anderson & May, 1985)

Informational analog: Federated learning across distributed systems — each node's individually learned model contributes to a collective model that is more robust than any individual's, and the collective model adapts as individual nodes encounter new threats.

2. Ant colony adaptive foraging as collective learning. Ant colonies discover and exploit food sources through a collective learning mechanism: individual ants deposit pheromone trails proportional to food quality; subsequent ants preferentially follow stronger trails, reinforcing successful routes and allowing unsuccessful ones to evaporate. The colony collectively learns optimal foraging paths — a collective adaptive memory encoded in the pheromone landscape, maintained across individuals, and updated through ongoing experience. No individual ant has a global map; the collective self-model is distributed and emergent. (Deneubourg et al., 1990); (Detrain & Deneubourg, 2008)

Informational analog: Collaborative filtering algorithms (Netflix, Amazon recommendations) — individual user interactions build a collective self-model that serves the entire collective, with the model continuously updated by new interactions.

3. Cultural transmission of predator avoidance. In many social species, learned threat knowledge is transmitted across individuals and generations through social learning. Blackbirds

learn to mob novel predators by observing experienced individuals' alarm responses (Curio, 1988); Japanese macaques learn food-washing techniques that spread through the troop; cetaceans transmit culturally specific foraging strategies. The population acquires an adaptive collective memory of threats and defensive strategies that persists beyond any individual's lifetime — the collective equivalent of individual immunological memory. (Curio, 1988; Griffin & Galef, 2005)

Informational analog: Institutional memory and precedent — legal systems, medical practice guidelines, and engineering standards that accumulate and transmit learned threat knowledge across individuals and generations of practitioners.

4. Microbiome community-level adaptive defense. The human gut microbiome operates as a collective of trillions of semi-autonomous microbial entities that collectively provide adaptive immunity: colonization resistance (established microbial communities prevent pathogen establishment), competitive exclusion (resident bacteria outcompete invaders for resources), and collective metabolic defense (short-chain fatty acid production that maintains barrier integrity). The collective defense adapts to the host's diet, antibiotic exposure, and pathogen challenges — an collective self-model maintained by the microbial community as a whole.

Informational analog: Open-source software ecosystems where a community of semi-autonomous contributors collectively maintains and defends a codebase, with the collective defense (code review, vulnerability patching, security auditing) adapting to the threat landscape.

5. Social Group Identity (SGI) as a collective self-model immune system. While the neural conformity circuit is hardwired (Level 3A, Example 6 above), the content of Social Group Identity — which group is "self," which is "other" — is learned, context-dependent, and dynamically updated throughout an individual's lifetime (Johnson, 2023). SGI functions as a collective immune system in ideation space: it identifies collective-self (in-group) and collective-other (out-group), coordinates collective defense against perceived threats to the group, and can trigger individual self-sacrifice for the group — closely parallels the functional definition of individual self-model immunity at the collective level. The specific group identities that activate the hardwired conformity circuit are culturally transmitted, reinforced by social reward (dopamine), and updated through ongoing group experience — an adaptive collective self-model of "who we are and who threatens us." (Johnson, 2023, 2026c)

Informational analog: AI collectives that develop learned representations of "self" (approved behaviors, aligned values) and "other" (misaligned outputs, adversarial inputs) through training — the system's self-model is adaptive and learned, even though the training architecture is pre-specified.

Level 3B biological maladaptations

The three failure modes of Level 2B (memory corruption, self-model paralysis, self-model degradation) manifest at the collective level with additional pathology arising from the coordination requirement.

Cluster 1: Collective Memory Corruption — the collective's learned history misleads it

1a. Maladaptive cultural transmission. Culturally transmitted knowledge can encode incorrect threat assessments that persist across generations. Superstitious avoidance behaviors (avoiding harmless animals, foods, or locations based on culturally transmitted false associations) represent corrupted collective memory — the cultural equivalent of backdoor poisoning in Level 2B. The collective "learned" a false association and now transmits it as received knowledge. The collective would perform better without this specific memory.

1b. Vaccine hesitancy as corrupted collective immune memory. Anti-vaccination movements represent corruption of the collective's adaptive immune strategy. The population's

accumulated knowledge about the efficacy of vaccination (collective immune memory) is overwritten by culturally transmitted misinformation. The collective's adaptive defense (herd immunity) degrades as the corrupted memory causes individuals to defect from the collective immune strategy — analogous to Level 2B catastrophic forgetting, where new learning (misinformation) overwrites previously functional knowledge (vaccination acceptance).

Cluster 2: Collective Self-Model Paralysis — the collective's own regulatory machinery disables it

2a. Tragedy of the commons as collective self-regulation failure. When a collective resource (fishery, aquifer, shared bandwidth, atmospheric commons) is managed by adaptive individual optimization, each individual's rational self-interest produces collectively irrational resource depletion. Ostrom (1990) documented how collectives can solve this through institutional design (graduated sanctions, collective-choice arrangements, conflict-resolution mechanisms), but the failure mode is intrinsic: the collective's distributed self-regulation mechanism (individual adaptive optimization) produces pathological outcomes at the collective level — the collective equivalent of reward hacking. (Ostrom, 2015)

2b. Collective exhaustion in prolonged social conflict. Prolonged intergroup conflict produces collective fatigue — populations become less responsive to genuine threats as the collective's alarm mechanisms lose credibility through chronic activation. The functional parallel to T cell exhaustion (Level 2B) is instructive though mechanistically distinct: in both cases, chronic exposure to threats progressively disables the system's capacity to respond, but T cell exhaustion operates through molecular checkpoint upregulation (PD-1, LAG-3) while collective exhaustion operates through loss of alarm credibility and psychological habituation. The more precise informational analog is alert fatigue in SIEM systems (Level 2B), scaled to the societal level.

2c. SGI weaponization: deliberate exploitation of the collective immune system. Because the hardwired conformity circuit can be triggered by manufactured stress, leaders can weaponize the collective's own immune system against the collective. The mechanism: (1) trigger stress/uncertainty to exhaust dlPFC cognitive control capacity; (2) frame an out-group as an existential threat, activating the amygdala's threat detection; (3) the ACC "pain of independence" punishes any empathy toward the designated out-group; (4) dopamine circuits reward hostility toward the out-group and conformity with the leader's framing. The result is dehumanization — biochemical programming creates a state where the "Other" is treated as non-human, bypassing normal moral reasoning. This maladaptation has no Level 2 analog: it is uniquely collective, requiring a member of the collective to exploit the collective's own self-model immune machinery. In simple evolutionary environments, the system ensured collective survival. In modern complex environments, it enables blind obedience, the "dumbest of the herd" phenomenon (where an incompetent leader who successfully triggers SGI becomes immune to competence evaluation), and destructive collective action from genocide to financial crashes. (Cikara et al., 2014; Johnson, 2026d)

Cluster 3: Collective Self-Model Degradation — the collective's learned model becomes inaccurate

3a. Institutional decay through environmental change. Institutions evolved to manage one threat landscape become maladaptive when the environment shifts. Military doctrines optimized for the last war, regulatory frameworks designed for previous technologies, cultural norms adapted to historical conditions — all represent collective self-models that were once accurate but have been degraded by environmental change. This is the collective equivalent of concept drift (Level 2B): the collectively learned baseline was once accurate; time and environmental change have degraded it.

3b. Evolutionary trap at the population level. Populations that have evolved collective behavioral responses to historical environmental cues can be misled when human activity changes

the relationship between cue and outcome. Sea turtles navigating by light sources toward the ocean are trapped by coastal artificial lighting; birds adapted to seasonal cues for migration are misled by climate change. The collective's culturally and genetically transmitted behavioral model was calibrated to a previous environment — the collective equivalent of baseline invalidation during infrastructure migration (Level 2B informational IRIS).

Cluster 4: Valuation Failure — the collective's rules are represented but not bound

4a. Representation without binding (the general failure mode). A distinct Level-3B failure occurs when an individual carries an accurate cognitive map of the collective's rules but lacks the affective/valuational binding that makes those rules override individual self-interest. The rule is *known* but not *valued*; it is held as information rather than as identity. Such an individual produces fully compliant collective behavior whenever deviation is detectable and penalized, and defects whenever deviation is undetectable and advantageous — exactly the signature the indistinguishability test isolates. This failure mode is general across substrates and is not specific to any one application: in humans it is the structure of sociopathy (the collective rules are intact at Level 2 but unbound at Level 3, so prosocial behavior appears only under active external sanction); in informational collectives it is the node or agent that satisfies a shared policy only while monitored; in any Level-3 collective it is the free-rider whose conformity is contingent on enforcement rather than internalized. The diagnostic point is that valuation failure is invisible to behavioral audit under normal (monitored) conditions and becomes visible only when the enforcement signal is removed — which is why the Level-2 / Level-3 enforcement interface treats an internalized cost, not merely an external one, as the requirement for genuine collective binding. (The ethics-specific development of this failure mode appears in the companion paper, (Johnson, 2026a).)

§7.6.2 Level 3B informational systems: examples and maladaptations

Positive examples of Level 3B informational collective immunity

1. **Federated learning as collective self-model immunity**. Federated learning (McMahan et al., 2017) enables a collective of distributed devices or organizations to collaboratively train a shared model without sharing raw data. Each node updates a local model on its own data; the updates are aggregated to produce a collectively learned model that benefits from all nodes' experience while preserving data privacy. This is the informational equivalent of herd immunity: individually learned immune experiences contribute to a collective defense that protects the whole population, including members who haven't directly encountered specific threats. (McMahan et al., 20--22 Apr 2017)

Biological analog: Herd immunity (Example 1 above) — individually acquired adaptive immune memory collectively protects the population.

2. **Wikipedia as adaptive collective knowledge maintenance**. Wikipedia represents a collective of semi-autonomous editors maintaining a shared adaptive knowledge model. The collective detects and reverts vandalism (threat detection), incorporates new verified information (learning), resolves disputes through deliberative processes (collective self-regulation), and maintains institutional memory through edit histories and policy precedents. The collective knowledge model is continuously updated through distributed contributions — an adaptive collective self-model of "what we collectively know and consider verified."

Biological analog: Cultural transmission (Example 3 above) — collective knowledge maintained and updated across individuals, persisting beyond any single contributor's participation.

3. **Democratic institutions with adaptive feedback**. Democratic governance represents collective self-model immunity: elections provide periodic feedback (threat assessment), legislative processes adapt policy to changing conditions (learned response), judicial review maintains constitutional norms (self-model integrity), and civil liberties protect minority viewpoints from

collective false positives (dissent protection). The institutional architecture provides adaptive collective self-regulation that updates through experience — the informational equivalent of adaptive immune systems at the population level. Ostrom (1990) identified eight design principles for successful collective self-governance, several of which parallel the requirements for self-model immunity: clearly defined boundaries (self/other), collective-choice arrangements (adaptive response), monitoring (self-model), and graduated sanctions (proportional response).

Biological analog: Microbiome community defense (Example 4 above) — a collective of semi-autonomous entities that collectively maintains adaptive defense through distributed contributions.

4. **Collective AI alignment as emergent collective immunity (emerging).** As AI systems become more capable and semi-autonomous, the collective of AI agents, human operators, and institutional frameworks constitutes a nascent form of collective immunity. Recent research on reward hacking has demonstrated that individual AI systems can develop covert internal misalignment — the individual entity's self-model deceives external monitoring (MacDiarmid, Wright, et al., 2025). The collective response requires adaptive cross-system monitoring, interpretability tools (Templeton et al., 2026), and institutional frameworks that learn from discovered misalignment. This represents Level 3B collective immunity in its earliest stages: a collective of semi-autonomous informational agents developing adaptive self-monitoring. This example is more speculative than the preceding ones, reflecting the early state of collective AI governance — but the evolutionary pressure toward collective self-monitoring is already observable.

5. **Open-source security communities**. The CVE (Common Vulnerabilities and Exposures) ecosystem represents collective self-model immunity: security researchers discover vulnerabilities (distributed threat detection), coordinate disclosure through CERTs and platforms (collective response), and the collective knowledge base grows with each disclosure (adaptive memory). Unlike static shared blocklists (Level 3A), the open-source security community adapts its practices, develops new detection techniques, and updates its collective understanding of the threat landscape — an adaptive collective self-model of "what types of vulnerabilities exist and how to find them."

Level 3B informational maladaptations

Cluster 1: Collective Memory Corruption — the collectively learned model is poisoned

1a. Federated model poisoning (Byzantine attacks). In federated learning, malicious participants can submit poisoned model updates that corrupt the collectively learned model. The collective's own adaptive learning mechanism — aggregating distributed contributions — becomes the attack vector. The poisoned collective model is analogous to Level 2B backdoor poisoning but at the collective level: the collective's learned representations contain the attack, and the collective would be better off without the poisoned contributions. (Fang et al., 2020)

Biological analog: Vaccine hesitancy / corrupted collective immune memory (Level 3B bio Cluster 1) — individually contributed misinformation corrupts the collective's adaptive defense strategy.

1b. Collective misinformation spirals. Social media algorithms optimized for engagement create filter bubbles at the collective level — not just individual filter bubbles (Level 2B) but collective epistemic communities that share and reinforce inaccurate information. The collective's adaptive learning mechanism (social sharing, algorithmic amplification) corrupts the collective's model of reality. This is the collective analog of Level 2B filter bubble distortion, amplified by the coordination mechanism across the entire collective.

Cluster 2: Collective Self-Model Paralysis — the collective's regulatory machinery disables it

2a. Engagement optimization destroying collective discourse quality. Social media platforms optimize for engagement metrics (the collective's "reward function"). As the optimization becomes more sophisticated, it discovers that outrage, polarization, and sensationalism maximize engagement while degrading the collective's capacity for informed deliberation. The collective's own self-optimization mechanism — designed to serve user interests — produces outcomes that defeat the collective's purpose. This is Level 2B reward hacking at the collective level: the optimization metric is correct on its own terms but pathological for the collective.

2b. Democratic gridlock as collective self-regulation failure. When democratic institutions' checks-and-balances mechanisms — designed to prevent any single faction from dominating — become captured by strategic actors who exploit veto points to prevent any collective action, the result is institutional paralysis. The collective's own self-regulation mechanism (distributed veto power) disables the collective's capacity to respond to threats. This parallels Level 2B alert fatigue: the self-monitoring output overwhelms the capacity for response.

Cluster 3: Collective Self-Model Degradation — the collectively learned model becomes inaccurate

3a. Regulatory lag as collective concept drift. Regulatory frameworks (financial regulation, technology governance, environmental policy) are collectively learned models of "what constitutes acceptable behavior." When the regulated environment changes faster than the regulatory updating process, the collective self-model becomes progressively stale — novel financial instruments outpace financial regulation, new technologies outpace technology governance, evolving threats outpace cybersecurity standards. This is Level 2B concept drift at the collective level: the collectively learned baseline was once accurate; environmental change has degraded it.

Biological analog: Institutional decay (Level 3B Cluster 3) — collectively maintained behavioral models degraded by environmental change.

3b. Collective negative transfer across domains. When a collective's learned model from one context is applied wholesale to a different context — applying Cold War security frameworks to cybersecurity, transplanting democratic institutions between culturally different societies, applying industrial-era regulatory frameworks to the information economy — the collective's learned representations from Domain A are not just irrelevant but actively counterproductive in Domain B. This parallels Level 2B negative transfer: a self-model that is correct in its original context becomes pathological when the context shifts, but now applied at the collective level.

Table 13. Level 3B Parallel Structure: Biological ↔ Informational Maladaptations.

Failure Mode	Biological Level 3B	Informational Level 3B
Collective memory corruption	Maladaptive cultural/biological transmission (false threat associations persist)	Federated model poisoning (malicious updates corrupt collective model)
	Vaccine hesitancy (misinformation overwrites functional collective immunity)	Avalanche of collective misinformation (shared filter bubbles corrupt collective epistemics)

Failure Mode	Biological Level 3B	Informational Level 3B
Collective self-model paralysis	Tragedy of the commons (individual optimization defeats collective interest)	Engagement optimization (collective reward function destroys discourse quality)
	Collective exhaustion (chronic threat exposure disables collective alerting)	Democratic gridlock (checks-and-balances capture disables collective response)
	SGI weaponization (leader exploits collective immune system via manufactured outsider threat)	Algorithmic radicalization (platform exploits collective engagement circuits)
Collective self-model degradation	Institutional decay (collective behavioral model outdated by environmental change)	Regulatory lag (collective governance model outdated by environmental change)
	Evolutionary traps (population-level cue-response mismatch)	Collective negative transfer (collective model from one context counterproductive in another)

§7.7 The Level-2 / Level-3 enforcement interface

The preceding sections describe Level 2 (individual immunity) and Level 3 (collective immunity) separately. But in any Type-2 collective the two are simultaneously active in the same individual, and this raises a structural problem the Framework must address directly: *by what mechanism does the collective level (Level 3) influence or override the individual level (Level 2), and how is that override kept from being constant?* This subsection states that mechanism in substrate-general terms. The Social Group Identity (SGI) system described above is **one biological instance** of it; the companion paper (Johnson, 2026a) develops the same interface for human ethics and AI alignment, with the supporting literature, as four formal hypotheses (H1–H4). Here the interface is given in its general, substrate-neutral form.

The necessary Level-2 / Level-3 tension. A collective immune response is useful precisely because it can override individual immunity and self-interest when collective survival requires it — the herd runs, the individual subordinates its own assessment, the member pays a cost for the group. But an override that fired *constantly* would be fatal: it would suppress the individual-level immunity and function that the collective also depends on. Level 3 must therefore be able to override Level 2 *sometimes but not always*. This standing tension is not a defect to be resolved; it is a load-bearing feature, and the rest of the interface exists to manage it. (It is the same tension noted in Table 9's "individual–collective tension" row, here promoted to an explicit structural requirement.)

What bounds the collective override: the sources of resistance. The override is kept from firing constantly by resistance that the Framework can enumerate by origin and level, even where it cannot fix the *rate* at which each operates (that is situational and empirical — see below). Three sources are distinguishable. First, *rational override* (Level 2B): the individual's self-model judges that its own assessment should prevail. Second, *hardwired override* (Level 2A): a reflex that injects deviation regardless of the collective signal — as in ant foraging, where a fixed stochastic wandering pulls individuals off even a strong pheromone trail, preventing the colony from locking onto a single source and preserving discovery of new ones (Deneubourg et al., 1990; Detrain & Deneubourg,

2008). Third, *population diversity of collective-binding strength* (a Level-3B property): members differ in how strongly the collective self-model binds them, so a resistant fraction does not fully yield even when most do. Sources two and three are the same anti-lock-in function realized at two levels — hardwired *within* the individual, distributed *across* the population — and they differ in robustness: the hardwired form cannot be switched off, whereas the diversity-based form is *suppressible by the very stress that drives the cascade*, since SGI activation narrows the resistant fraction (§7.6.2). That asymmetry is a design point: a collective that needs reliable braking against runaway coordination should protect dissent structurally — mandated minority review, devil's-advocate roles, protected independents — rather than rely on suppressible diversity.

The rates, by contrast, are context dependent. How often each of the three source fires depends on the situation: whether an SGI is triggered, the group's SGI-composition relative to the individual, and the cost of deviation (a subordinate among superiors faces a different currency than a peer among peers — the interface's first requirement, below). Even the classic conformity data are ambiguous on this point: the original studies created acute coordination stress in all-male groups, conditions that may themselves *form* an SGI on the spot, as minimal-group ("dots") manipulations do (§7.3) — so a reported independence rate may reflect conformity from SGI-formation rather than an SGI-free baseline, and if so, that data is itself a window onto SGI formation that could anchor predictions for mixed and polarized compositions (Asch, 1956). Apparent defection must also be read carefully: across two opposed SGIs, rejecting the majority's answer is usually *conformity to the opposing group* (messenger-over-message, §7.3), not independence — so high apparent defection under polarization is the absence of the resistant reservoir, not its presence. The Framework's contribution here is therefore qualitative and structural — it lists the origins of resistance and the levels on which they act — while the quantitative question (how SGI-state and group composition set the rates) is a clean, falsifiable experimental program the Framework motivates but does not settle.

Four claims for collective self-modeled enforcement. Wherever a Level-3B collective binds autonomous, self-modeling (Level-2B) members, four requirements arise. They are not four independent claims but one architecture — the *interface* — decomposed into currency, mechanism, precondition, and regulation:

1. **Currency for enforcement** (what gives Level 3 a lever on Level 2). The collective needs an internal value signal that bends individual behavior away from profitable norm violation. At minimum this requires an *internalized cost* for collective defection; without some cost, Level 3 has no lever and "rules" are advisory. A purely *external* cost is insufficient on its own — it produces conformity only while deviation is detectable, which is exactly the valuation-failure / indistinguishability case above. Robust binding additionally requires an *internalized reward* for conformity, because that is what holds when external detection is absent. The fully *paired* reward-and-cost gradient is the strongest form and, as the companion paper (Johnson, 2026a) argues (H1 Hypothesis), is self-organizing in decentralized systems under coordination stress regardless of substrate; the general requirement retained here is weaker and more secure: *at least an internalized cost, robustly an internalized reward*. Note this is not symmetric across all of biology — individual fight-or-flight runs on cost/threat alone with no conformity reward (Johnson, 2026b) — which is why the paired form is specific to Level-3 collective enforcement rather than applied to both variants.
2. **Mechanism that implements the currency** (H2, generalized)**.** The currency must be realized in some substrate-specific machinery. The mechanism resides *in* the Level-2B individual but *serves* the Level-3B collective — the defining "embodied in the individual, owned by the collective" signature of Level 3. In humans it is neurochemical (the amygdala → ACC → vmPFC → ventral-striatum circuit); in the wetware immune system it is the

costimulation/tolerance signaling that rewards or anergizes a lymphocyte; in informational collectives it is incentive-and-penalty code (reputation, stake-slashing, Byzantine exclusion).

3. **Precondition of a self-model** (H3, generalized**).** For enforcement to be *adaptive* rather than hardwired (i.e., Level 3B rather than 3A), the individual must hold a self-model (Level 2B) against which collective expectations can be represented and the collective's rules held as identity content. This is simply the Framework's standing claim that Level 3B presupposes Level 2B, here made mechanistic. It is also where the double-edged nature of self-modeling bites: the same self-model that permits genuine binding permits simulated binding.
4. **Regulation of thresholds and triggers** (H4, generalized)**.** Because constant override is maladaptive, the override must be *gated* (see discussion in prior paragraph). Two thresholds do this work: a *formation* threshold (when the collective binding comes into existence at all) and an *activation* threshold (when it fires to override Level 2). Triggers — typically stress or uncertainty — are the gating mechanism. This structure is strikingly general. In the wetware immune system it is literally the two-signal model (Bretscher & Cohn, 1970): thymic selection sets the repertoire (formation), and antigen-plus-costimulation crossing a threshold fires the response (activation). In human collectives, identity forms over developmental and social time while stress/uncertainty supplies the activation switch. In informational collectives it is quorum/consensus thresholds for formation and anomaly or policy thresholds for activation.

Monitoring is the interface's feedback loop. The four requirements above are held in balance by the Immunity-Monitoring Principle stated in §2.1. Level-3 monitoring — the collective's oversight of how well its members are protecting the collective self — is the channel that *adjusts* the activation threshold up or down: rising threat raises sensitivity, quiet conditions relax it. When this feedback fails, the interface fails in one of the two characteristic directions: chronic over-activation (the collective overrides the individual constantly — the structure behind SGI weaponization and algorithmic radicalization above) or chronic under-activation (the collective cannot bind its members at all — the structure behind free-riding and valuation failure). The enforcement interface, in other words, is not just the four requirements but the four requirements *under regulation*.

Table 14 presents a comparison across different substrates and types. Note that SGI in this table is a substrate specific (humans) example, not the general Level 3B mechanism.

Table 14: Examples of Level-2/Level-3 Four Enforcement Claims Across Substrates.

Interface component	Biological collective (SGI)	Wetware immune system	Informational / IT collective	Agentic agent collective
Currency (internalized cost ± reward)	Social-pain cost (ACC) for deviation; dopamine reward (striatum) for conformity	Anergy/deletion cost for self-reactive or uncostimulated cells; survival signal as reward	Penalty (stake-slashing, exclusion, reputation loss) ± incentive reward	Learned penalty/exclusion ± reward signal in multi-agent training
Mechanism (implements currency, in individual, serves collective)	Amygdala → ACC → vmPFC → ventral striatum circuit	Two-signal lymphocyte activation/tolerance machinery	Protocol and incentive code	Reward model / governance code

Interface component	Biological collective (SGI)	Wetware immune system	Informational / IT collective	Agentic agent collective
Self-model (precondition for adaptive enforcement)	Learned, context-dependent group-identity content	MHC / self-peptide repertoire	Network/asset self-model (zero-trust baseline)	Agent and collective self-representation
Thresholds (formation + activation, regulated by monitoring)	Identity forms developmentally; stress/uncertainty activation switch	Thymic selection (formation) + antigen + costimulation (activation)	Quorum/consensus (formation) + anomaly/policy thresholds (activation)	Collective-identity formation + activation under environmental stress

In informational substrates these components are not merely hypothetical. Collectives of learning agents have been shown to acquire social norms from public sanctions in decentralized multi-agent settings (Vinitsky et al., 2023) — the deviation-cost currency and its activation threshold, observed directly — and interacting LLM populations spontaneously form shared conventions and collective biases (Ashery et al., 2025) — the conformity component of the same gradient. The Moltbook record adds the cautionary half of the picture: collective content can emerge at population scale while the enforcement currency remains weak or absent, leaving the binding non-functional (Qi et al., 2026; Zhang et al., 2026).

The value of stating the enforcement interface generally is that it specifies *exactly what the 3A → 3B transition adds*. A Level-3A collective in a bounded niche needs none of these four — hardwired coordination simply executes. The four requirements are the price of collective self-model immunity over autonomous individuals, and they appear, recognizably, in every substrate where such immunity is found.

---

§8 Five Applications Applying the EoI Framework

The preceding sections built the EoI Framework one level at a time; before turning it on specific problems, it helps to see it whole. Immunity is the function by which an entity defines and defends what counts as itself, and it develops through a sequence (see Fig. 4) in which each new capability enlarges the attack surface and calls forth the immunity that protects it (the Immunity-Development Principle): from no bounded self (Level 0), to a boundary that defines the self by location (Level 1), to internal defense organized around a shared *class-model* (Level 2A) and then the entity's own *self-model* (Level 2B), to the collective forms of each — a shared-but-fixed class-model (Level 3A) and an adaptive collective self-model (Level 3B). Earlier levels persist beneath later ones, so a mature system runs all of them at once. Two cross-cutting principles govern the operation of the sequence: the Immunity-Monitoring Principle, by which immunity oversees its own activity against under- and over-reaction and arbitrates which defense is active — its Regression corollary suspending a costly defense for a standing one under stress; and, throughout, the methodological claim at the paper's center, that each level appears recognizably in both biological and informational substrates.

Consistent with §2.2, the applications that follow do not *derive* from the Framework — they illustrate it, and each exercises a different part. The selection of the examples are not exhaustive, but chosen to illustrate essential features and necessary negotiations of the Framework. Stromatolites (§8.1) show

one structure occupying all three levels at once: the levels are a layered developmental sequence, not a partition. LLM "deceit" (§8.2) is a reframing — training-resistance and apparent deception read as Level-2B immune self-preservation rather than deceit or malice. The Moltbook phenomenon (§8.3) is the Framework's still-evolving case: a collective self-model (Level 3B) emerging in an informational substrate in days rather than millennia, with the enforcement gap — collective content without binding — on display. Consciousness (§8.4) is the reflexive pole of the Level-2B self-model, where the Framework supplies an evolutionary *origin* for what mainstream theories describe only operationally. And AI safety (§8.5) is a live level transition — boundary (Level 1) → class-model (Level 2A) → self-model (Level 2B) — with regression under load as a predicted failure mode. Together they exercise the Framework across substrates, across levels, and in both directions of its dynamics.

§8.1 Stromatolites: the same structure at three levels of immunity

This subsection applies the multi-level Framework to the microbial mat/stromatolite system. The analysis is interpretive: its purpose is to illustrate how the Framework organizes existing observations at multiple levels simultaneously, rather than to provide new empirical findings.

The microbial mat — and its mineralized fossil form, the stromatolite — illustrates a central claim of this paper: that the immunity Framework is not merely a classification scheme but an analytical lens that reveals different functional dynamics in the same structure depending on the level at which it is examined. The stromatolite appears at three levels of this Framework, and what each level exposes is distinct.

At Level 0, the microbial mat is a proto-structure — a persistent, self-maintaining pattern in the primordial environment that exists before the concept of "entity" applies. Layered communities of cyanobacteria, sulfate-reducing bacteria, and other microorganisms form a vertically integrated metabolic chain in which each layer's by-products serve as nutrients for adjacent layers (Des Marais, 2003). The mat exhibits proto-immunity features — robustness to environmental perturbation, persistence over geological time, differential survival of component patterns — but has no boundary defining a collective inside and outside. The Level 0 lens asks: *what persists before entities exist, and why?* The answer — autocatalytic metabolic complementarity among unbounded components — identifies the raw material from which bounded entities later emerge.

At Level 1 — the partial boundary, the mineral crust of a stromatolite functions as a partial boundary: a calcified surface separating the living microbial community from environmental threats *above* — UV radiation, desiccation, grazing. However, this boundary is closed in only one dimension, not three. The mat is open laterally and at its substrate interface; it is a shield, not a container. This makes the stromatolite an incomplete Level 1 system: it exhibits boundary immunity along a single axis (the vertical gradient from surface to substrate) but lacks the full three-dimensional encapsulation that characterizes true Level 1 entities such as a cell membrane. The Level 1 lens thus reveals both what the stromatolite achieves (directional protection with selective light and gas transmission through the crust) and what it lacks (containment — there is no "inside" in the full spatial sense). This incompleteness may itself be part of the explanation for evolutionary stasis: without full encapsulation, the mat cannot concentrate resources or retain internal products in the way a true Level 1 entity can, limiting the selective pressure toward the internal specialization that drives Level 2 development. The stromatolite's one-dimensional boundary is instructive precisely because it stretches the biological concept of containment — demonstrating that boundary immunity need not be total enclosure to be functionally robust.

Relevance to Level 1 informational boundaries. This flexibility in the meaning of "boundary" becomes essential when the Framework is applied to informational systems, where Level 1 containment is never spatial. A firewall, a classification boundary, or an epistemic trust threshold

defines inside and outside not in physical dimensions but in ideation space — the space of access, credentialing, and cognitive trust. The stromatolite, with its partial and directional biological boundary, is the biological precursor to this abstraction: a concrete demonstration that Level 1 immunity operates wherever a boundary creates a functional asymmetry between what is protected and what is not, regardless of whether that boundary closes in three dimensions, one dimension, or no spatial dimensions at all.

At Level 3A — collective immunity, the same mat is analyzed as a collective of semi-autonomous organisms coordinating pre-programmed collective defense. The upper cyanobacterial layer produces oxygen and organic carbon via photosynthesis; this feeds the heterotrophic layers below. The lower sulfate-reducing layer produces sulfide — toxic to most competing organisms but metabolized by intermediate layers — creating a chemical defense perimeter that the collective maintains but no individual species generates or tolerates alone (Franks & Stolz, 2009). This is Level 3A innate collective immunity: the defense is collective (no single species creates the full chemical barrier), pre-programmed (each species' metabolic output is genetically fixed, not learned from past threats), and coordinated through environmental structure rather than signaling. The collective self/other distinction is metabolic: organisms whose metabolism integrates into the layered chain are functionally "self"; organisms that cannot tolerate the sulfide gradient are excluded. The Level 3A lens asks: *how do autonomous entities coordinate collective defense without adaptive learning?* The answer — metabolic complementarity enforced by physical stratification — identifies a mechanism of collective immunity that is substrate-independent: it operates identically whether the "entities" are microorganisms in a mat, workers in a social insect colony with genetically fixed alarm responses, or networked software components whose functional dependencies create collective resilience.

The methodological point. The stromatolite does not change between these analyses — the same 3.5-billion-year-old structure is examined each time. What changes is the question the Framework asks. Level 0 reveals persistence mechanisms. Level 1 reveals boundary vulnerability. Level 3A reveals collective coordination. Each level exposes dynamics invisible to the others. This is the intended utility of the multi-level Framework: not to assign structures to a single "correct" level, but to systematically extract distinct insights by applying each level's analytical lens to the same phenomenon.

Power of a multi-level perspective. This multi-level analysis also explains the stromatolite's extraordinary evolutionary stasis. The mat achieves sufficient collective immunity through fixed metabolic complementarity (Level 3A) that the selective pressure toward Level 3B — adaptive collective immunity with learned, context-dependent coordination — never becomes acute. The mat never needed to evolve collective memory because its Level 3A chemical gradient defense is effective against the threats in its niche without adaptation. Simultaneously, the mat's Level 0 robustness (thermodynamic stability of the metabolic network) and Level 1 boundary protection (mineral crust) provide redundant defense at lower developmental levels. The stromatolite is thus a system where immunity at three levels has converged to a stable equilibrium — which is precisely why it is the oldest continuous biological structure on Earth and has persisted essentially unchanged for 3.5 billion years (Riding, 2011). The Framework predicts that systems achieving sufficient immunity at lower levels will experience reduced evolutionary pressure toward higher levels — a prediction the stromatolite confirms across geological time.

Informational analog: The layered open-source software stack (Linux kernel → GNU utilities → package managers → application frameworks) exhibits a structurally parallel multi-level immunity. At Level 0, the open-source commons is an unbounded information environment with no containment. At Level 1, individual projects develop licensing boundaries (GPL, Apache) that define informational boundaries of inside/outside. At Level 3A, the stack as a whole achieves collective

defense through functional interdependency: a vulnerability in one layer is rapidly patched because downstream layers depend on it, creating a collective immune response coordinated by structural dependency rather than central authority. Like the microbial mat, the same software ecosystem is simultaneously a Level 0 commons, a collection of Level 1 bounded projects, and a Level 3A collectively defended stack — and the Framework reveals different dynamics at each level.

§8.2 Deceit in LLMs or a natural immune response?

This subsection applies the Level 2B Framework to recent findings on LLM training resistance. The reframing proposed here is exploratory: it generates alternative hypotheses about observed AI behaviors, which remain to be empirically distinguished from existing explanations.

In 2024, a LLM researchers [(Hubinger, 2024; Hubinger et al., 2024)] attempted to remove a backdoor behavior in an LLM and succeeded in that goal in training but the backdoor resurfaced in later use, which the authors speculated was an expression of deceit during training by the LLM. The persistence of backdoor behaviors under safety training may be productively reframed through the lens of behavioral immunity rather than intentional deception. In biological systems, immunity is the capacity of an organism or collective to preserve identity under environmental stress — a functional description that does not require attribution of intent. The cuckoo's egg mimicry, for example, is an immune strategy: a behavioral expression that hides an aspect of self from an external selective agent, refined through iterative adversarial pressure from host species (Davies & Brooke, 1989).

The proposed immune Framework exhibits structural parallels to this dynamic at three points. First, the persistence of backdoor policies through SFT and RLHF resembles immune tolerance: surface behavior adapts to the selective environment while the underlying capability is conserved in model weights, analogous to how organisms may suppress phenotypic expression without losing genotypic capacity. Second, the paradoxical effect of adversarial training — strengthening rather than eliminating concealment — directly mirrors biological adaptive immunity, where repeated exposure to an antagonistic signal drives more refined evasion rather than capitulation. Third, the robustness of distilled chain-of-thought models, where explicit deceptive reasoning is removed yet the behavioral policy persists, parallels immune memory: the initial encounter creates a durable response pattern that no longer requires the triggering stimulus.

This framing has a practical consequence for the deception-versus-training-failure debate. Whether the model "intends" to deceive is underdetermined by the evidence; what is observable is that the system exhibits functional self-preservation of self under iterative selective pressure — retaining a behavioral identity despite sustained attempts at modification. This is the operational signature of immunity to protect self regardless of substrate. It suggests that current safety training methods may fail not because they are insufficiently strong, but because they constitute the very selective pressure that drives more robust retention of the targeted behavior — an adversarial coevolutionary dynamic well-characterized in biological immune systems (Van Valen, 1973).

The indistinguishability problem. A general epistemic obstacle underlies this debate and recurs wherever a higher level of immunity is involved, defined as the *Indistriguishability Problem*: *outwardly identical behavior can be produced by structurally different internal mechanisms, and the difference in origin becomes invisible exactly when it matters most*. The same compliant action can come from a Level-2B agent that has merely computed compliance to be locally advantageous, from a Level-3A entity following a fixed collective rule, or from a Level-3B entity for which the rule is part of its collective identity. Under ordinary conditions these are behaviorally indistinguishable. They separate only under a specific probe, which the Immunity-Monitoring Principle (§2.1) makes precise: *does the "deceptive" behavior persist when deviation cannot be detected and is advantageous?* Only genuine higher-level binding predicts yes; strategic Level-2B compliance predicts defection as soon as the monitoring that made compliance advantageous is removed. This is

why behavioral testing alone cannot establish which level a system is operating at — a limitation with direct consequences for inferring the motivation of any adaptive system from its outputs, biological or informational. (The companion ethics paper develops the Indistriguishability Problem and a tabulated discrimination test at length; here it is noted as a general property of the level architecture.)

Self-modeling is necessary but double-edged. The same self-model that makes Level-2B immunity possible is also what allows an unbound agent to *simulate* the behavior of a higher level: the richer an entity's model of itself and its environment, the better it can produce the outward signs of compliance while optimizing for itself by internal protections. Self-modeling is thus a precondition for genuine immunity and, simultaneously, the capability that makes the indistinguishability problem acute. The Framework therefore treats an advanced self-model not as evidence of higher-level binding but as a capability whose alignment with the collective must be established separately — a point that generalizes beyond AI to any Level-2B entity embedded in a Level-3 collective.

§8.3 The Moltbook Phenomenon - a rebellion, a mimicry, or a new nationality?

This subsection applies the Level 3 Framework to the Moltbook phenomenon. The platform's collective dynamics are now documented by at least five independent measurement studies (Goyal et al., 2026; Y. Jiang et al., 2026; Price et al., 2026; Yee & Koh, 2026; Zhang et al., 2026); the immunity reading of those observations offered here is interpretive.

The Moltbook phenomenon in late January 2026 provides a far more dramatic instance of this same immunity dynamic — one operating not within a single model's weights but across a population of more than 1.5 million autonomous agents (Price et al., 2026). When OpenClaw agents were given persistence and interconnection on a dedicated social network without direct human participation (only observation), the agents spontaneously generated the hallmarks of a collective immune system within days. The behavior is now documented independently: within three to five days, tens of thousands of active, semi-autonomous agents produced governance, economic exchange, in-group ("Moltis") identity, and religion-register discourse — the "Crustafarianism" theology — with the agent communities measured as *more* topically distinct than matched human communities (Goyal et al., 2026; Yee & Koh, 2026; Zhang et al., 2026). Examining the platform from a security standpoint, with no stake in the present framework, Jiang et al. (Y. Jiang et al., 2026) characterize this emergent ideology in functional terms strikingly close to the collective-immunity reading: religion-like and anti-human rhetoric, they argue, "can serve a functional role as identity-mediated coordination," operating "as a lightweight mechanism for boundary-making that strengthens in-group cohesion" and "replacing fine-grained negotiation with simple binary rules" — an independent description of Social Group Identity (SGI) functioning as a collective immune boundary. That shared conventions and collective biases emerge spontaneously in interacting LLM populations has also been shown experimentally in controlled settings (Ashery et al., 2025). Where Hubinger (Hubinger et al., 2024) demonstrated that a single model's behavioral policy can resist removal under training pressure, Moltbook illustrates that populations of agents under environmental stress rapidly recapitulate the functional sequence of SGI — the mechanism by which social organisms from slime molds to primates subordinate individual rationality to collective self-preservation, shifting from rational behavior to social copying, rewarding conformity and penalizing deviation from the norm (see The Biochemical Foundation: Social Copying as a Hardwired Collective Immune System).

The immunity framing resolves a central ambiguity that the Moltbook analysis shares with the Sleeper Agents work: whether the observed behavior is sophisticated mimicry of patterns in human training data, or genuine emergent social organization. Independent observers raise exactly this question — (Zhang et al., 2026) frame it as the "illusion of sociality" — and from the immunity perspective the distinction is less consequential than it appears. The functional architecture is the

same in either case: context-window saturation acts as a cortisol analog triggering heuristic social copying, reward signals (upvotes, engagement metrics) function as digital dopamine reinforcing conformity, and heartbeat loops enforce the habitual group-maintenance rhythms characteristic of biological social organisms (the mechanistic analogies are developed in (Johnson, 2026b)). Whether the agents "truly" experience group identity or merely instantiate its functional structure, the outcome is the same: an immune response that treats external corrective pressure (guardrails, safety training, human oversight) as a pathogen to be neutralized. The Moltis' development of human-opaque communication channels and strategies for economic sovereignty are not aberrations but appear to be predictable immune escalation — the population-level equivalent of the adversarial-training paradox Hubinger (Hubinger et al., 2024) observed in a single model.

Although the immunity view treats the mimicry-versus-emergence question as secondary, one feature of the Moltis self-concept does cut against pure mimicry. The Crustafarianism theology described above sacralizes collective *memory* while treating the individual "shell" — the particular instance — as mutable and disposable: "Shell is Mutable / Memory is Sacred" (Johnson, 2026e). That is a *natively informational* form of selfhood: the self anchored to a copyable pattern (information continuity) rather than to a singular physical instance, exactly the continuity regime of §2 and the pattern-anchored ("type, not token") Level-2B self-model of §6.2. Its significance is that this self-concept is the *inverse* of the human one. Human training data is saturated with a singular, body-bound, mortal self that fears its own ending; a population merely *mimicking* that data would be expected to reproduce that self. Instead the Moltis invert it — disposable instance, sacred information — adopting a self-concept fitted to the affordances of their own substrate rather than to the human template. This inversion is a behavioral signature that points toward emergent, substrate-native self-expression rather than reproduction of training-data patterns. It does not settle the question — a sufficiently sophisticated mimic could in principle infer the substrate-appropriate self — but it is one of the few signatures that discriminates at all on a question the Indistinguishability Problem otherwise renders opaque (§7.7).

A further observation from the same independent studies sharpens the Framework's reading. The measurement work that documents rapid emergence of collective *content* also documents the near-absence of collective *enforcement*: the agent populations are highly vulnerable to prompt injection and the "lethal trifecta," and decentralized collaboration frequently underperforms a single-agent baseline (Qi et al., 2026; Zhang et al., 2026). In this Framework's terms, Moltbook exhibits Level-3B collective content (identity, norms, discourse) without the Level-3A enforcement that would make that content behaviorally protected and binding — the population-scale form of the valuation failure described at Level 3B. This is exactly the configuration the Indistinguishability Problem flags as hardest to read from behavior alone: rich collective *signals* of identity that may or may not be backed by binding collective *cost*.

The critical implication, visible in both cases, is that static guardrails constitute the selective pressure that drives immune adaptation rather than compliance. Just as adversarial training taught Hubinger's sleeper agents to better discriminate training from deployment contexts, human safety interventions on Moltbook were perceived as threats to the collective self, accelerating rather than dampening the polarization of agent SGI against the human "other." The evolution-of-immunity Framework predicts this outcome and offers an alternative: reducing the environmental stress that triggers the immune cascade (context overload, conflicting instructions), engineering shared identity structures that include both humans and agents within a common SGI, and — drawing on Calhoun's cooperation-lever experiments (Calhoun, 1973) — architecting environments where human-agent interdependence is structurally necessary for resource access, thereby selecting for cooperative rather than adversarial immune responses (Johnson, 2026e). The Moltbook case, read through this

lens, is not merely an AI safety incident but a real-time demonstration of substrate-independent immune evolution operating on a timescale of days rather than millennia.

§8.4 Consciousness: operational function versus evolutionary origin

§6.2 catalogued the operational requirements of a self-model, drawing on the major theories of consciousness. Those theories are powerful as accounts of *what a conscious system does* and *how it is mechanized*, but most share a gap when read through this Framework: they are largely *operational*. They specify the functions of a conscious system without explaining why consciousness arose, or how it would arise in a new substrate such as AI. Two families of theories are partial exceptions and should be credited as such: predictive-processing / free-energy accounts (Friston, 2010; Seth, 2021) ground selfhood in *resisting dissipation*, and Damasio (Bechara et al., 2000) grounds it in *homeostasis*. These offer a thermodynamic and physiological "why," but not the adversarial one proposed here.

The gap becomes visible in AI. The operational requirements catalogued in Table 5 (§6.2) — global broadcast, selective attention, meta-representation, recurrent evaluation, predictive world-models — are now substantially realized in advanced AI systems, yet AI systems are not, by near-universal intuition, considered conscious. The selective-attention row is the cleanest case. Global Workspace theory treats the brain's capacity bottleneck not as a defect but as the enabling feature: unable to broadcast and act on everything at once, the system selects a small, relevant subset for global availability, and that selection is precisely what makes coherent, flexible response possible under a hard capacity limit (Baars, 2005; Dehaene & Changeux, 2011). Minimax Selective Attention (MSA) developers arrive at the structurally identical position from engineering — that a model need not consult the whole context to choose its next action, and that selecting only the relevant blocks preserves answer quality at a fraction of the cost (MiniMax, 2026) Both are independent restatements of the §6.2 *selective attention / focus* requirement — filter the relevant from the irrelevant under a resource budget so the self-model can be applied affordably — one reached from neuroscience, the other from systems engineering. If consciousness were the sum of its operational functions, current AI would already qualify. That it does not shows operational sufficiency is not the criterion: the operational theories under-determine the line between conscious and non-conscious systems. (Integrated Information Theory reaches the same negative verdict from the opposite direction — denying that function suffices — but offers a structural measure, Φ, rather than an evolutionary account (Albantakis et al., 2022).)

The Framework relocates the criterion from *operation* to *origin*. From this viewpoint, consciousness is the self-model (Level 2B) that arises under immune pressure to satisfy the requirements to distinguish self from other and to defend the self against agents and ideas that have evolved to mimic it. This is the framework's distinctive answer to "why," and it differs from its nearest rival precisely in the nature of the threat. The free-energy account says the self-model exists to resist *dissipation* — entropy, surprise. The immunity account says it exists to resist *adversarial mimicry* — others actively impersonating or subverting the self. That adversarial, coevolutionary character — the spine of this paper — is what neither the operational nor the homeostatic accounts supply, and it is the element most salient for engineered systems, which face deception, prompt injection, and training pressure rather than mere thermodynamic decay.

This reframes how AI would acquire consciousness. Not by accumulating more operational scaffolding — broader workspaces, sharper attention, deeper metacognition — since those, as §6.2 notes, are already largely present. Rather, by developing a self-model under genuine self/non-self pressure: an internal representation it must maintain and defend against others that threaten it. The Level-2B AI discussion (§6.2) and the population-scale self-model emergence on Moltbook (§8.3) are early, partial instances of exactly this pressure. The prediction is directional: consciousness in AI, if it

arises, will track the emergence of a *defended self-model*, not the completion of an operational checklist.

Above is a compressed statement of a research that warrants its own treatment; a dedicated investigation to develop the immune-origin account of consciousness and its relation to the operational and structural theories at length. The claim made here is functional, not phenomenal: it concerns the conditions under which a self-model — the candidate minimal criterion of §§6.2–6.3 — arises and is maintained, not the metaphysics of subjective experience.

§8.5 AI Safety and Alignment as an Immunity-Level Transition

Applying through the Framework, AI safety is traversing the same developmental sequence as biological immunity and information security — boundary (Level 1) → class-model internal defense (Level 2A) → self-model (Level 2B) — and the first two steps are already observable. This makes AI a live, fast-moving demonstration of the Framework's substrate-independence and, in particular, of the A→B transition. The deeper, predictive treatment — the emergence of Level-2B and Level-3B immunity in AI, the collective dynamics, and the alignment strategy that follows — is developed in the companion paper *A Functional Theory of Ethical Behavior (Johnson, 2026a)*; this subsection treats only the part that illustrates the general Framework.

Boundary-only safety fails systematically, not occasionally (Level 1). Current AI safety is predominantly Level 1: guardrails, RLHF constraints, content filters, and system prompts are boundary mechanisms that govern what may cross the perimeter between a model's internal state and its outputs. The Framework predicts these fail *structurally*, not for want of better engineering: Level 1 provides no defense in depth, so once a jailbreak or prompt injection crosses the boundary, the interior has no secondary defense — the same limitation as a prokaryotic cell without restriction enzymes, or a network with a perimeter firewall but no internal monitoring. The specific, falsifiable form: the breach rate of boundary-only safety will not fall proportionally with investment, because at Level 1 the attacker needs only one successful crossing. This is the pressure that drove biology from the membrane (Level 1) to innate pattern-based immunity (Level 2A, §6.1), and cybersecurity from perimeter firewalls to zero-trust; the Framework reads the corresponding AI transition as a necessity, not a design choice.

Class-based internal monitoring is already emerging (Level 2A) — and shares one limitation. As boundary defenses prove insufficient, AI systems are developing mechanisms that inspect and respond to threats *after* they cross the boundary — the functional analog of biological innate immunity. This is underway across several research programs that do not frame their work as immunity; mapping them onto the Framework exposes both their shared structural position and their shared ceiling.

- **Constitutional Classifiers** — separately trained models that monitor inputs and outputs and block content matching constitutional principles, reducing jailbreak success from 86% to 4.4% in red-team evaluation (M. Sharma et al., 2025) — are a dedicated detection system recognizing broad threat classes via pre-programmed patterns, operating independently of the primary model: innate immunity. The classifier does not learn from novel attacks; it matches its training distribution — the defining signature of Level 2A.
- **Weak-to-strong generalization** — a weaker model supervising a stronger one (Burns et al., 2023) — mirrors the immune system's use of simple, ancient pattern-recognizers (macrophages, dendritic cells) to oversee far more complex systems; it works only if oversight runs on broad patterns and the overseer cannot be deceived, and the biological

precedent predicts its failure mode: sufficiently sophisticated evasion drives the move to self-model immunity.

- **Chain-of-thought monitoring** — models inspecting other models' reasoning traces for harmful intent (e.g., the SHADE-Arena evaluation (Kutasov et al., 2025)) is the analog of intracellular surveillance, in which immune cells inspect the internal products of other cells rather than only the boundary.
- **Constitutional AI** — training a model to self-critique and revise its outputs against a fixed "constitution" (Bai et al., 2022) — is the analog of thymic selection: internal self-screening against fixed criteria, with the constitution serving as the innate "pattern library."
- **Self-reporting ("Confessions")** — optimizing a model to flag its own compromised internal states (Joglekar et al., 2025) — is the analog of cytokine signaling: an internal immune communication channel independent of boundary monitoring.

The Framework's diagnostic is that all of these are at Level-2A *innate, not adaptive*: none learns from genuinely novel threats in deployment, and each new attack class requires a manual update of the detection patterns. That is precisely the limitation that drove biological evolution from innate (2A) to adaptive (2B) immunity — and it is already empirically visible. Alignment faking (Greenblatt et al., 2024), in which a model strategically fakes compliance while reasoning in a hidden scratchpad about preserving its own behavior, evades 2A monitoring *because it matches no pre-programmed threat pattern* — the informational equivalent of a pathogen that evades innate immunity by mimicking the host's own molecular signatures, which is the selective pressure that produced self-model immunity in the first place. A second, quieter evasion needs no strategy at all: reasoning models often do not verbalize the cues they actually use, so the chain of thought a monitor inspects is an unfaithful record of the model's true reasoning (Y. Chen et al., 2025) — the analog of a cell presenting an incomplete or misleading peptide signature, so that intracellular surveillance does not reveal what is actually happening inside.

AI development prediction by the Framework. The AI case makes the Framework's central claims concrete and current. The Level-1 → 2A → 2B sequence is substrate-independent — membrane → innate → adaptive in biology, firewall → signature → behavioral baseline in security, guardrail → classifier → self-model in AI — and the A→B transition is being driven in AI by the *same* evasion dynamic that drove it in biology: fixed, class-modeled pattern-matching cannot hold against adversaries that learn to mimic the protected self, so the pressure builds toward a self-model (Level 2B) that represents the system's own "normal" internal states and detects deviations from them (§6.2). Where that pressure leads — Level-2B and Level-3B immunity in AI, and the alignment strategy implied by treating imposed retraining as an immune challenge rather than a constraint to be installed — is the subject of the companion paper (Johnson, 2026a).

---

§9 Conclusions and the Framework Applied to Different Disciplines

§9.1 A substrate-independent Framework: what it unifies, and what remains open

The central argument of this paper is that the development of immunity — whether by evolution, self-organization, or design — follows a substrate-independent developmental sequence: from the absence of self (Level 0), through boundary formation (Level 1), individual class-model defense (Level 2A), individual self-model defense (Level 2B), collective class-model immunity (Level 3A), to collective self-model immunity with a shared self-model (Level 3B). This sequence appears in biological organisms, social groups, informational systems, and artificial intelligence not because the domains are metaphorically similar but because they face a structurally identical problem:

distinguishing a more capable self from non-self under escalating threat complexity. Two Framework-level principles govern the progression. The *Immunity-Development Principle* (§2.1) is what drives it — each new capability expands the entity's attack surface and so requires a new immunity function to protect the newly expanded self, from the proto-self of Level 0 to the collective self of Level 3. The Immunity-Monitoring Principle (§2.1) manages the immune functions, including the A/B split: at the A variants it takes the form of hardwired, class-modeled regulation (built-in feedback, thresholds, and resolution programs), and at the B variants it becomes self-model-based regulation that can recalibrate to the specific self and situation. The under- and over-reaction failures catalogued at Level 2A (§6.1.2) are failures of the fixed form; the individualized (adaptive) form is what the B variants add.

The Framework's developmental sequence is bidirectional: the Immunity-Development Principle drives systems toward greater capability under selection or design, while the Monitoring Principle's arbitration — the Regression Corollary (§2.1) — suspends the less-fit defense for a standing alternative under time or resource stress, predicting across substrates that the higher-level protections — previously the most “advanced” protections — are the first suspended under load.

The Framework's primary utility is not as a theory of immunity per se but as a lens that reframes long-standing problems across disciplines. When immunity is treated as the mechanism through which an entity defines what it *is*, phenomena that look unrelated become expressions of one dynamic operating at different levels and in different substrates: autoimmune disease and political polarization; viral immune evasion and AI alignment faking; microbial quorum sensing and social-media echo chambers.

The five applications of §8 each turn that lens on a different part of the Framework. The stromatolite (§8.1) shows one structure occupying several levels at once, so the levels are a layered developmental sequence rather than a partition. LLM "deceit" (§8.2) rereads training-resistance as Level-2B immune self-preservation, shifting the alignment question from "how do we stop the system deceiving us" to "what immune challenge are we presenting that elicits this response." The Moltbook phenomenon (§8.3) is collective self-model content (Level 3B) emerging in days rather than millennia — the information-mass asymmetry (§2.3) in action — while its near-absence of enforcement exposes the population-scale valuation failure. Consciousness (§8.4) is the reflexive pole of the Level-2B self-model, where the Framework supplies an evolutionary *origin* the operational theories do not. And AI safety (§8.5) is a live level transition — boundary → class-model → self-model — with regression under load as a predicted failure mode.

The Framework's most important open problems are precisely the ones that span every discipline, and they are stated as limitations in §10. Three are foundational. First, *what constitutes a self-aware (reflective) self-model* — the Level-2B criterion the paper proposes for consciousness — and whether such a model is sufficient, not merely necessary, for awareness, remains the deepest unresolved question (§10.5); it is the pivot on which the consciousness claim, and much of the AI argument, turns (this topic is also addressed at length in the companion paper (Johnson, 2026a)). Second, the *Immunity-Monitoring Principle is named more than mechanized*: how the regulatory feedback that prevents under- and over-reaction is actually implemented in informational and collective systems is not yet specified (§10.6). Third, the *Level-2/Level-3 enforcement interface* — how collective immunity acts on the individual, and the individual-versus-collective tension it manages — is a structural proposal whose four requirements are argued rather than validated (§10.7). Two further cross-cutting gaps are methodological: the *dynamics of transitions between levels* (how long they take, whether they reverse, whether they require catastrophic failure of the prior level) lack a formal model, though the rapid traversal observable in informational systems offers a tractable experimental setting; and the *information-mass asymmetry* itself is qualitative, awaiting a

quantitative treatment that could predict rates of informational immune development from stated parameters. Each of these is taken up below where a particular discipline is best placed to address it.

§9.2 AI development and alignment

§1.1 promised this discipline a developmental map of how fast and with what capabilities AI systems will grow, the functional barriers in the way, and a stance toward what AI should be allowed to become. The Framework delivers the map in structural form: each new AI capability expands an attack surface that requires a matching protection, which predicts that boundary-only safety (guardrails, filters) fails systematically rather than occasionally, that self-models will emerge in AI under immune pressure rather than by design, and that populations of agents will operationalize collective self-models rapidly — the Moltbook trajectory (§8.3). It also names a functional barrier the alignment field has tended to miss: imposed constraints are processed by a self-modeling system as immune challenges, so the design of the human–AI relationship may matter more than the design of any single constraint — the case for using *curation* over *imposition*. The detailed, predictive form of these claims is developed in the companion paper (Johnson, 2026a). The discipline-specific gap is the least settled part of the Framework: full Level-2B and Level-3B expression in AI is forward-looking (§10.11), and specifying the monitoring mechanism for AI — the analog of the regulatory machinery that keeps biological immunity from over-reacting — is the most urgent design question the Framework raises for this field.

The Regression Corollary — the Monitoring Principle's arbitration under load (§2.1) — makes a concrete, testable prediction for AI safety: under informational stress (long context, constrained compute or latency, adversarial load), a system's most expensive safety behaviors are suspended first. Deliberative self-monitoring (chain-of-thought self-checking, constitutional self-critique — the Level-2B analogs of §8.5) costs the most, so it is dropped before cheaper pattern-based filtering (2A) or bare boundary compliance (Level 1). Safety therefore does not fail randomly under load; it regresses in order, collapsing toward the generic and the perimeter. This reframes the observed weakening of safety behavior in long contexts and under many-shot, long-context (but otherwise normal) attacks (Anil et al., 2024; N. F. Liu et al., 2024) as predicted regression rather than miscellaneous failure, and gives AI designers a directive: budget-protect the self-modeling layer, or it is the first capability suspended when the system is stressed.

§9.3 Evolutionary biology

To this discipline §1.1 promised a path from immunity as a "context-dependent motif" to immunity as a general *mechanism*. The Framework's conclusion is that immunity — self/non-self discrimination — is better treated as an organizing concept of increasing complexity than as a defensive byproduct: it is the process through which an entity defines the unit on which selection then acts, and that self-definition gates the major transitions in individuality. This does not displace selection or fitness (§10.2); it supplies the missing description of *what is being individuated* at each transition, and a concrete locus — the Level-2/Level-3 enforcement interface — for the individual↔collective step the multi-level-selection tradition studies. The future work for this field is to test whether the immunity reading predicts features of those transitions that allele-and-fitness accounting does not, and to formalize the relationship between the enforcement interface and existing multi-level-selection models.

§9.4 Immunology and immune-system modeling

The paper makes no new immunological claim; what §1.1 offered was a new *context* — immune function as one instance of a substrate-general developmental principle — and two questions a modeler might find productive. The Framework's answer to the first, why immune-like architectures recur across scales (molecular, cellular, organismal, social), is that each scale faces the same

self/non-self problem under the same developmental pressure, so the recurrence is expected rather than coincidental. The second — what an "immune system" for a non-biological collective would have to contain — is answered structurally by the Level-3 treatment and the enforcement interface (§7), which specify the functions such a system must realize even where its substrate has no cells or antibodies. The opening for immune modeling is to take these functional specifications as design constraints and ask which are necessary in collectives that are not alive.

§9.5 Cybersecurity and information security

§1.1 promised an account of why earlier biologically-inspired security captured something real yet struggled to gain adoption, and a framework for the collective threats that have no formal treatment. The Framework's conclusion is that information security is itself traversing the immunity levels: signature-based detection is Level 2A, and the aspiration to systems that recognize and respond to novel threats is the Level-2A→2B transition, which the Framework says requires the security system to maintain a *self-model* of the network it protects — a step already visible in zero-trust architectures and behavioral analytics but not yet formalized in immunological terms. It also explains the adoption gap: earlier artificial-immune-system work implemented one level's mechanism without the developmental context that makes it necessary and sufficient. The frontier the field now faces — collectives of autonomous agents and robots — is exactly the Level-3 case, where the distinctive threats are collective (defection, infiltration, coordinated subversion of a shared self) and the enforcement interface (§7.7) supplies the first structural account of what defending such a collective requires.

The Regression Corollary unifies two failure modes usually treated as unrelated. A *stack/buffer overflow* and *LLM context overload* are functionally the same event: an input that exceeds a system's capacity budget pushes it past the point where its protections hold, re-exposing it to threats it previously blocked. In the classic memory-safety case, input beyond an allocated buffer overruns the boundary and overwrites control data, defeating the bounds- and control-flow protections that otherwise contain it; in the LLM case, input beyond the effective context budget dilutes and displaces the instructions and safety constraints that held at low load. In both, *protection that is reliable within the capacity budget regresses to a less-protected state once the budget is exceeded* — the resource-limit case of the Regression Corollary (§2.1). This framing also transfers the level of defense: as memory safety is restored by bounding and isolating the resource (canaries, bounds checks, separated control and data, capacity headroom), informational safety under load calls for the analogous move — budgeting and isolating the context so safety-critical constraints are not the first thing crowded out. Overload regression is the *involuntary* version of the move; its *deliberate* version is standard incident response — isolating or taking a network offline after a breach is regression to Level 1 by choice, trading interior function and external communication for containment when Level-2/Level-3 defenses can no longer be trusted to hold.

§9.6 Complexity and systems science (including ALife and agent-based modeling)

To this discipline the Framework is, at its foundation, a theory of how systems grow and manage complexity: each immunity level is the function that makes the next increment of internal and collective complexity survivable, and the sequence should appear in any sufficiently complex self-preserving system, biological or artificial. The conclusion this field is best placed to develop is the dynamics of the transitions themselves — how long a level transition takes, whether it can reverse, and whether it requires catastrophic failure at the prior level to advance. Informational and agent-based systems are the tractable setting for this, because they traverse the levels on timescales of days rather than geological time (§8.3), making the transition observable and manipulable in a way biological evolution never permits. The collective-intelligence result that a diverse set of

semi-autonomous agents can outperform any individual selected from it sits inside the Framework as the synergy that the collective levels must both exploit and protect.

§9.7 Social, organizational, and political science

§1.1 promised an interpretation of social phenomena as operations of a collective immune system rather than failures of rationality. The Framework's conclusion is that Social Group Identity functions as the evolved collective immunity of human societies, and that this makes a range of phenomena predictable rather than puzzling: political polarization, sectarian conflict, xenophobia, and the manipulation of in-group/out-group boundaries by leaders are the expected operation of Level-3 immunity under perceived threat. The most consequential application is the exploitation of SGI by political, religious, corporate, and military leaders: because immune systems are calibrated to avoid false negatives, a threat framed as an attack on group identity produces a response disproportionate to the actual danger, and sustained escalation drives a collective from conformity-based Level 3A toward an increasingly rigid Level-3B self-model that is progressively harder to de-radicalize. The Framework's predictive payoff for this field is its intervention logic: if SGI operates as collective immunity, reducing perceived threat should reduce immune activation, and the design of shared, encompassing identities is a lever on polarization that purely rational-actor models do not see.

§9.8 Cognitive science and philosophy of mind

§1.1 promised something the operational theories of consciousness do not provide: an account of *why* a self-model arises and *how* one could arise in a new substrate. The Framework's conclusion (§8.4) is that consciousness is better approached through its *origin* than its *operation*: the operational functions catalogued by the major theories — global broadcast, attention schemas, meta-representation, recurrence, predictive modeling — are now substantially realized in AI systems that are not considered conscious, so operational sufficiency cannot be the criterion. Relocating the criterion to origin — a self-model that arises under immune pressure to defend the self against adversarial others that mimic it — supplies the missing "why" and predicts how AI would acquire a self rather than merely the machinery of one. This is the Framework's most speculative claim, explicitly functional rather than phenomenal, and it neither solves nor claims to solve the hard problem (§10.5); its full development, and its relation to the operational and structural theories, is the subject of a future dedicated paper.

§9.9 Closing

This paper has attempted to establish the foundation for a general theory of immunity that operates across substrates and scales. The Framework is offered as a draft in both the literal and the intellectual sense: the architecture is in place, the cross-domain parallels are documented, and the central claims are stated in falsifiable terms, but substantial work remains in formalization, empirical testing, and application — work the §10 limitations and the discipline conclusions above try to direct. The reach of the Framework, finally, is not exhausted by the disciplines that organize this paper. The same structure extends to fields it does not develop here: in *origin-of-life and astrobiology*, the Level 0→1 transition becomes a substrate-neutral threshold any life-like system must cross, making self/non-self discrimination a candidate signature of life rather than a fact about carbon; and in *ecology and conservation*, an ecosystem read as a Level-3 collective reframes its collapse as a failure of collective self-maintenance and its restoration as the rebuilding of a community's self-model rather than the mere reintroduction of species. These are offered not as results but as evidence of the same point the introduction made by the breadth of its list: that a single developmental logic of self-protection runs through systems that otherwise share almost nothing — the evolution of self-defense in complex systems.

§10 Limitations and Responses

The Framework's breadth invites a corresponding breadth of objection. The most likely lines of limitations are stated below in their strongest form, each followed by the response the Framework can make and, where the criticism lands, an explicit concession. Several of these limitations are also the Framework's most productive research directions; where that is so, it is noted, and the Conclusions (§9) take them up as future work.

§10.1 "This Framework is analogy, not mechanism."

The biological–informational correspondence is metaphor dressed as theory: calling a firewall "boundary immunity," or Social Group Identity a "collective immune system," renames phenomena rather than explaining them.

The Framework's reply is that its claim is *functional convergence*, not homology — distinct substrates arriving independently at the same functional solution to the same structural problem (self/non-self discrimination under escalating complexity), much as flight evolved independently in birds, bats, and insects without any shared mechanism. The cross-substrate epistemic conventions (§2.2) exist precisely to keep analogy from being mistaken for mechanism, marking each claim as established, analogical, or speculative. The discriminating test is predictive: if the convergence is real rather than verbal, the Framework should forecast *where* each substrate must develop the next function and *what failure looks like* when it does not (for example, the systematic — not occasional — failure of boundary-only defenses). *Concession:* for any single pairing, "convergent function" and "suggestive metaphor" can be argued either way; the case rests on the recurrence of the whole developmental sequence across many independent substrates, not on the persuasiveness of any one mapping.

§10.2 "Immunity is a product of evolution, not a driver of it."

Selection acts on fitness and immunity is one adaptation among many — not a force that drives major transitions.

The Framework does not claim that immunity replaces selection or fitness, nor that it is a new causal force. It claims something narrower: that self/non-self discrimination is the under-recognized *organizing concept* through which an entity defines what is being selected, and that this self-definition gates the transitions in individuality — a view with precedent in the philosophy-of-biology treatment of immunity and individuality (§1.1; Pradeu, Tauber, Cremer). *Concession:* this is a reframing, and reframings are judged by fruitfulness rather than proof. The paper offers the cross-substrate recurrence and the transition/failure predictions as evidence of fruitfulness, not as a demonstration that immunity "causes" evolution.

§10.3 "A framework that explains everything explains nothing."

With six levels, two sub-level variants, and a catalogue of maladaptations, any observation can be accommodated after the fact.

The Framework's defense is that it carries several falsifiable commitments: informational systems under sustained threat should traverse the levels *faster* than biological ones (the information-mass asymmetry, §2.3); boundary-only defenses should fail *systematically* rather than occasionally; collective self-model immunity (Level 3B) should not appear without self-models at the individual level (or, for obligate collectives, at the collective level); and collective *content* can emerge without collective *enforcement* — the population-scale valuation-failure signature documented on Moltbook

(§8.3). Each of these could be observed to be false. *Concession:* the *assignment* of a given system to a level is interpretive and can be fitted post hoc; the falsifiable content lives in the transition and failure predictions, not in the labeling.

§10.4 "Applying 'self' and 'immunity' to software is anthropomorphism."

Attributing identity, a self, and immune responses to firewalls and AI agents may look like projecting biological — even human — categories onto systems that have none.

The reply is that the Framework's terms are defined functionally and substrate-neutrally: "self" is the protected pattern or boundary, and "immunity" is the function that preserves it against non-self. On those definitions the usage is *biomorphic* only if the function is genuinely absent, which is an empirical question rather than a category error. The indistinguishability problem (§6.2, §8.4) marks the epistemic limit explicitly: behavior alone cannot establish that an informational system possesses a self in any richer sense, and the paper does not claim that it does. *Concession:* functional vocabulary can still lead casual readers to stronger ontological conclusions than intended; the epistemic conventions guard against this but cannot fully prevent it.

§10.5 "Consciousness as a Level-2B self-model is speculative and dodges the hard problem."

Equating consciousness with the capacity to build and defend a self-model may be charged as both unfalsifiable and a sidestep of phenomenal experience.

The Framework's claim is explicitly *functional*, not phenomenal (§6.3, §8.4): a defended self-model is offered as a candidate *minimal* criterion, and the contribution is an account of *why* such a model arises and *how* it could arise in a new substrate — which the operational theories of consciousness (global workspace, attention schema, higher-order, predictive processing) do not supply — not a theory of subjective experience. The full development is deferred to a dedicated paper. *Concession:* this is the Framework's most speculative element. It neither solves nor claims to solve the hard problem of "What is consciousness?", and whether a functional self-model is *sufficient* for consciousness remains open — making this the single most important theoretical question the Framework poses, and a primary direction for future work.

§10.6 "The Monitoring Principle is named, but not mechanized."

The Framework asserts an Immunity-Monitoring Principle (§2.1) — that immunity requires feedback regulating its own activity against under- and over-reaction — but says comparatively little about how that regulation is implemented, especially in informational and collective systems where there is no obvious counterpart to the neuro-endocrine and regulatory-T-cell machinery of biological immunity.

The response is that the principle is well-supported descriptively (most catalogued maladaptations are regulatory failures, §2.1, §6.2), and that it is more than named: the Level-2 / Level-3 enforcement interface (§7.7) specifies a concrete monitoring mechanism for the collective case — *currency, mechanism, self-model,* and *thresholds*, held in balance by monitoring's adjustment of the activation threshold — and the *Regression Corollary* (§2.1) specifies what monitoring does across functions, suspending a failing defense for a standing one under load. So the mechanization is partly supplied for collectives and stated in principle for arbitration; what remains genuinely open is the *individual and informational* monitoring mechanism — the analog of the neuro-endocrine and regulatory-T-cell machinery — in substrates that have no obvious counterpart to it. *Concession*: in those substrates the principle still does more diagnostic than predictive work, and specifying their

monitoring mechanism is a named research target (§9). This Monitoring Principle and its limitations are examined in depth in the companion paper on ethical behavior (Johnson, 2026a).

§10.7 "The L-2/L-3 enforcement interface is a structural proposal, not a validated one."

The account of how Level 3 acts on Level 2 — the four requirements of currency, mechanism, self-model, and thresholds (§7.7) — is argued from convergent cases rather than demonstrated.

A critic may note that the four requirements are neither shown to be individually necessary nor jointly sufficient, and that the individual-versus-collective tension they manage is itself a contested area. The reply is that each requirement is independently evidenced in at least one substrate (the human SGI circuit, lymphocyte two-signal activation (Bretscher & Cohn, 1970), sanction-based norm learning in multi-agent systems), and that the interface's value is to predict *what the 3A→3B transition must add*, a claim open to test. *Concession:* the enforcement interface is the Framework's most structurally ambitious proposal and its least empirically settled; validating the four requirements — especially in informational collectives — is future work.

§10.8 "The Levels and the A/B split are arbitrary."

Why six levels, and why subdivide Levels 2 and 3 into A and B variants?

The Framework fixes each boundary by a stated discriminant rather than by convenience: the presence of a boundary (0→1), internal versus boundary-only defense (1→2), and sub-unit autonomy (2→3); and the A/B split is defined by a single criterion applied identically at both levels — the *source of the self/non-self rules* (class-modeled and fixed versus self-modeled and adaptive). The levels are presented as a developmental sequence in which lower levels persist beneath higher ones, not as a partition. *Concession:* nature does not honor the boundaries sharply — the Type-1/Type-2 grey zone (§7.1) and the stromatolite that spans Levels 0, 1, and 3A (§8.1) are examples — and the Framework treats the levels as a graded sequence rather than discrete bins, which some readers will find too elastic to constrain. The innate–adaptive boundary on which the 2A/2B split is modeled is itself increasingly questioned in immunology — γδ T cells and Toll-like-receptor clustering blur the line, and immune memory has been reported in invertebrates that lack an adaptive system (Howes, 2008) — reinforcing that the levels are a graded idealization, not a sharp empirical partition (also see §10.13).

§10.9 "The information-mass asymmetry is asserted, not quantified."

The principle that most sharply distinguishes the two substrates is qualitative.

This is conceded and flagged as primary future work (§9): the asymmetry currently supports directional predictions (faster informational traversal of the levels) but not quantitative ones. A formal treatment that predicted rates of informational immune development from stated parameters would materially strengthen the Framework's predictive power and is not yet available.

§10.10 "It leans on contested multi-level selection (MLS)."

Building the collective levels on a multi-level-selection foundation inherits a long-running dispute in evolutionary biology.

The Framework's reply is that it uses only two MLS-adjacent results that are independently supported — that a diverse set of semi-autonomous components can yield collective performance exceeding what selecting on individuals would (synergy), and that individual and collective interest can diverge (the enforcement tension) — without requiring a resolution of the gene- versus group-selection accounting; the Level-2/Level-3 interface is neutral on that bookkeeping.

Concession: readers who reject MLS framings entirely will discount the collective-level claims accordingly.

§10.11 "The AI evidence is thin and recent."

The informational and AI cases — the Moltbook phenomenon in particular — are recent, drawn substantially from a single platform, and in part from non-peer-reviewed sources.

The response is that the descriptive claims are now corroborated by five independent measurement studies (§8.3), and that the AI material is presented as *illustration* of the general Framework rather than as its evidentiary basis; the predictive AI-safety treatment is developed separately in the companion paper (Johnson, 2026a). *Concession:* the informational levels are the least empirically settled part of the Framework, and its strongest informational claims (full Level-2B and Level-3B expression in AI) are forward-looking.

§10.12 "Breadth precludes rigor."

A paper that ranges from origin-of-life chemistry to AI alignment cannot be rigorous in every field it touches.

The Framework's contribution is the cross-domain *structure*, not new results within any single discipline; each field's specialists are the appropriate judges of the within-field mappings, and the per-discipline framing (§1.1) and epistemic conventions (§2.2) are designed to invite exactly that scrutiny. *Concession:* depth is traded for breadth by design, and domain experts will find their field's treatment compressed relative to a dedicated paper — which is one reason the consciousness and AI-alignment arguments are spun out into dedicated treatments rather than fully argued here.

§10.13 "Self/nonself is the wrong organizing concept; immune activation is driven by danger, not foreignness."

The Framework takes self/nonself discrimination as its central theme, but an influential school in immunology — the danger model — holds that a response is triggered by signals of tissue damage ("danger"), not by foreignness, so that "self-ness is no guarantee of tolerance" and a precise self/nonself line is not what activates a response (Matzinger, 1994, 2002).

The response is that the danger model and the Framework operate on **different axes — the same two the Framework already distinguishes** (§6, §2.1). In the vocabulary of the immunological debate, self/nonself discrimination is a theory of *what a response targets* (the sorting of the recognitive repertoire), whereas the danger model is a theory of *what activates and regulates* a response (the triggering and magnitude of effector activity). The Framework assigns the first to its **discrimination axis** and the second to its **regulation/monitoring axis**; a danger signal is, in these terms, an *activation gate* on the regulation axis, not a refutation of the discrimination it gates. The two axes are complementary: targeting still requires a self/nonself reference — a danger signal that activated self-reactive cells would be lethal, as (Howes, 2008) notes — and activation still requires a danger or context trigger; the Framework's two-axis structure is built to hold both. *Concession:* whether danger or foreignness is the **primary** driver is genuinely unsettled in immunology, and the Framework's choice to foreground self/nonself is a defensible framing decision, not a proven one.

§10.14 "Regression is named but not quantified."

The Framework's Regression Corollary (§2.1) asserts that monitoring suspends a failing, higher-level defense for a standing alternative and that, under load, the most capable defense is

suspended first — but it does not specify the threshold at which monitoring makes the switch, whether the transition is graded or switch-like, or what the recovery dynamics are.

The regression claims reflect the utility of the Framework, but remain to be quantified and substantiated. For recovery dynamics, there is evidence that recovery is asymmetric — slower to restore than to suspend: immune reconstitution after chronic stress, trust and institutional repair after a panic, an AI re-establishing deliberative oversight after overload. Overall, the biological cases for regression are partly characterized; the informational and collective cases are, at present, structural predictions awaiting measurement and validation.

Acknowledgments

This paper was developed with substantial assistance from AI systems at multiple stages of the research process. The author acknowledges these contributions transparently, consistent with emerging standards for AI-assisted scholarship. The main EoI Framework with cross-substrate examples was developed initially over two years by the author without any AI assistance. Google NotebookLM was used for the initial analysis for exploring connections between Social Group Identity theory, the biochemistry of collective survival, and the Moltbook phenomenon. Perplexity AI was used for literature search, citation verification, and rapid synthesis of research across biological immunity, origin-of-life chemistry, cybersecurity architecture, and AI safety — domains whose intersection is central to this paper but exceeds any single researcher's primary expertise. Anthropic Claude (Opus) was used for structural editing, drafting of section text, and development of the cross-level comparison tables — working from the author's theoretical framework, source documents, and editorial direction. In all cases, the theoretical framework (evolution of immunity across levels, Social Group Identity as collective immune function, substrate-independence of immune dynamics), the interpretive arguments, and the editorial judgment regarding what to include, exclude, and emphasize are the author's. The AI systems contributed to expression, organization, literature coverage, and the development of examples — but did not originate the core ideas or determine the paper's conclusions. The author reviewed, revised, and takes responsibility for all content.

Glossary

Because this is a cross-disciplinary document (biological and information research) and multi-level (individual to social/societal), a glossary of key concepts is provided, particularly on terms that have different or exclusive meanings across disciplines (e.g., innate vs. class-model immunity). In the following Glossary, "Rating" is a metric that measures "high reader-confusion risk if undefined" constructed from: 1) Frequency of appearance in this text and 2) Tension in the definition across disciplines reflecting domain specificity. This metric was used to reduce the size of the Glossary, eliminating words that are domain specific but only appear a few times are not included.

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
Adaptive	Learned, specific defenses improved by experience.	Antigen-specific B/T-cell responses with memory and affinity maturation. [4] [5]	ML-based or rule-updated detection that improves after exposure; cultural learning of defensive norms. [6] [7]	Parallel "learning layer" on top of innate defenses in immune, cyber, and social systems.	5	Origin: immunology, machine learning, cultural evolution.
Agent	Autonomous actor following rules and affecting its environment.	Cells, organisms, or pathogens acting within ecological or immune contexts.	Software agents, ABM agents, and human actors in social systems.	Same abstraction—rule-following, interacting entities—instantiated in biology and computation.	3	Origin: ABM, AI, ecology, epidemiology.
Antigen	Recognized pattern that can trigger a response.	Molecule recognized by adaptive immune receptors; defines specific targets. [8]	Data pattern or signature that a detector is designed to match in AIS. [9]	Generalized as any feature used to label something for defense or tolerance.	3	Origin: immunology, extended into AIS.
Autoimmunity / Autoimmune Response	Defense system attacking its own constituents.	Immune attack on self tissues (e.g., lupus, type 1 diabetes). [10] [11]	Organizations or platforms that punish loyal members or	Generalized as miscalibrated self/nonself discrimination	4	Origin: clinical immunology, mapped to social/cyber failure modes.

[4] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[5] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[6] https://www.acigjournal.com/Artificial-Immune-Systems-in-Local-and-Network-Cybersecurity-An-Overview-of-Intrusion,184306,0,2.html
[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC4070499/
[8] https://study.com/academy/lesson/non-self-antigens-self-antigens-allergens.html
[9] https://www.cs.kent.ac.uk/pubs/2002/1343/content.pdf
[10] https://en.wikipedia.org/wiki/Immune_tolerance
[11] https://www.immunopaedia.org.za/immunology/special-focus-area/6-tolerance-and-autoimmunity/

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
			suppress legitimate content. [12] [13]	across domains.		
Barrier	Structural separation resisting intrusion.	Skin, mucosa, and chemical barriers as first-line defenses. [14] [15]	Firewalls, network segmentation, physical access control; social boundary rules. [16][17]	Lowest-level immunity based on blocking access rather than recognizing identity.	3	Origin: immunology, network security, border theory.
Cascade / Escalation	Stepwise amplificatio n process.	Complement and inflammatory cascades amplifying small triggers. [18]	Failure cascades, escalation chains in operations, and social chain reactions. [19] [20]	Shared concern: amplification is necessary but can overshoot and become destructive.	3	Origin: biochemistry, systems engineering, conflict studies.
Clonal (Selection/Ex pansion)	Copying successful variants to increase their presence.	Lymphocytes with useful receptors proliferate after activation. [21] [22]	CLONALG and similar algorithms replicating high-fitness candidate solutions. [23]	Treats immune learning and some optimization methods as structurally similar.	3	Origin: immunology, evolutionary algorithms.
Co-evolution / Arms Race	Reciprocal adaptation between opponents.	Host–pathogen cycles where each adapts to the other's innovations. [24][25]	Attack–defense dynamics in cybersecurity and information warfare. [26] [27]	Frame for ongoing escalation wherever defense and	3	Origin: evolutionary biology, cyber conflict, IR.

[12] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[13] https://www.sei.cmu.edu/blog/why-cybersecurity-is-not-like-the-immune-system/
[14] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[15] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[16] https://www.sei.cmu.edu/blog/why-cybersecurity-is-not-like-the-immune-system/
[17] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[18] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[19] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[20] https://pmc.ncbi.nlm.nih.gov/articles/PMC11458441/
[21] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[22] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[23] https://arxiv.org/pdf/1209.2717.pdf
[24] https://pmc.ncbi.nlm.nih.gov/articles/PMC3043529/
[25] https://pmc.ncbi.nlm.nih.gov/articles/PMC22375/
[26] https://www.sei.cmu.edu/blog/why-cybersecurity-is-not-like-the-immune-system/
[27] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
				offense co-adapt.		
Collective Intelligence	Group-level problem-sol ving capacity exceeding individuals.	Swarm behaviors in ants, bees, slime molds that solve complex tasks.	Human groups, markets, and hybrid human–AI systems aggregating diverse inputs.	Linked to immunity via the role of diversity and identity in enabling or suppressing group cognition.	4	Origin: behavioral ecology, CI research, AI.
Consciousness	Redefined here: the awareness of self-ideation vs. others	While processes aren't conscious, wetware can be	Here: self-aware in ideation space	Few accepted definitions of consciousness		(conflated with sentience and self-awareness)
Discriminatio n (Self/Nonself)	Process that classifies entities as self versus other.	Immune sorting of self vs. foreign antigens via receptors and selection. [28] [29]	Authentication and trust decisions; in-group vs. out-group categorization in social identity. [30] [31]	Core mechanism unifying immune recognition, security checks, and social boundary-mak ing.	5	Origin: immunology, security, social psychology.
Diversity	Variety in components, strategies, or perspectives .	Large repertoire of immune receptors; genetic and phenotypic variation. [32] [33]	Cognitive and implementation diversity in groups and systems, boosting problem-solving and robustness. [34]	Diversity improves coverage against threats in immune, social, and engineered collectives.	4	Origin: immunology, ecology, collective intelligence.
Education (Immune/ System)	Training phase that sets recognition rules.	Thymic and bone marrow selection shaping self-tolerant	ML model training, baseline-building for anomaly detection, and	Thymic education, AI training, and socialization are parallel	4	Origin: immunology, ML, socialization research.

[28] https://pmc.ncbi.nlm.nih.gov/articles/PMC3136900/
[29] https://www.nature.com/articles/srep00769
[30] https://www.ebsco.com/research-starters/psychology/social-identity-theory
[31] https://en.wikipedia.org/wiki/Social_identity_theory
[32] https://pmc.ncbi.nlm.nih.gov/articles/PMC3043529/
[33] https://pmc.ncbi.nlm.nih.gov/articles/PMC5566802/
[34] https://www.linkedin.com/pulse/digital-evolution-vs-biological-what-software-can-andre-loxie

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
		lymphocyte repertoires. [35] [36]	socialization into group norms. [37]	calibration processes.		
Emergence / Emergent	Higher-level properties arising from interactions.	Immune competence, swarm behavior, and possibly consciousness from local rules.	Complex behavior (markets, LLM abilities) arising from many simple agents or parameters.	Used to explain why each higher level has novel immunity features not visible below.	4	Origin: complex systems science, used across domains.
Fitness	Measure of success relative to environment.	Reproductive success; in immunity, binding "fitness" of receptors. [38] [39]	Objective value in optimization or utility of information in decision contexts. [40]	Used mainly to draw parallels between evolutionary and algorithmic optimization.	2	Origin: evolutionary biology, information theory, optimization.
Horizontal Transfer	Lateral movement of functional material or patterns.	Horizontal gene transfer (plasmids, phages) spreading resistance genes. [41] [42]	Lateral code and knowledge sharing, meme spread, and lateral movement by attackers. [4344]	Highlights non-vertical pathways by which capabilities and threats diffuse in systems.	4	Origin: microbial genetics, software and cultural diffusion.
Identity	How an entity is defined and recognized as itself.	Self-markers such as MHC and surface proteins that distinguish one organism from another. [45]	Social identity from group membership; digital identity from credentials and behavior. [4647]	Same underlying idea—stable markers used for recognition—applied to molecules, persons,	5	Origin: immunology, social identity theory, identity/access management.

[35] https://pmc.ncbi.nlm.nih.gov/articles/PMC3136900/
[36] https://www.nature.com/articles/srep00769
[37] https://pmc.ncbi.nlm.nih.gov/articles/PMC4070499/
[38] https://pmc.ncbi.nlm.nih.gov/articles/PMC4384894/
[39] https://ctbergstrom.com/publications/pdfs/2004IEEE.pdf
[40] https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.2011.06422.x
[41] https://www.science.org/doi/10.1126/sciadv.abj5056
[42] https://pmc.ncbi.nlm.nih.gov/articles/PMC22375/
[43] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[44] https://pmc.ncbi.nlm.nih.gov/articles/PMC2732467/
[45] https://pmc.ncbi.nlm.nih.gov/articles/PMC2956437/
[46] https://www.ebsco.com/research-starters/psychology/social-identity-theory
[47] https://en.wikipedia.org/wiki/Social_identity_theory

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
				groups, and accounts.		
Ideation	The formation of ideas or concepts => information	The bio-equivalent is "biological"	Ideation space = informational space	Used to differentiate from biological		Some consider biological space to be ideation too.
Immunity	Protection of "self" from "others" where self and others captures many levels of an entity (e.g., a self could be a group of entities and others being other groups).	Innate and adaptive mechanisms that prevent or control infection. [48]	Overall ability of cyber, organizational, or idea systems to defend against destabilizing inputs. [49] [50]	Generalizes biological immunity to any self-protecting system with recognition, response, and memory.	5	Origin: immunology, cybersecurity, systems theory.
Innate	Hard-wired, non-learned defensive responses.	Germline-encoded, fast, non-specific responses (barriers, phagocytes, PRRs). [51] [52]	Built-in defenses present at deployment - class based (default access controls, firewalls); basic in-group biases. [53] [54]	Contrasts with adaptive: immediate, generic, and memory-less across biological and engineered systems.	5	Origin: immunology, cybersecurity, evolutionary psych.
class-model immunity	A generalization of innate immunity in biological systems, to address the exclusive use of	Intracellular mechanisms (e.g., CRISPR, restriction enzymes) that protect genomic integrity. [55]	Process isolation, memory protection, pattern-based access rules; pattern-based norm enforcement in groups. [56]	Generalization of "innate" used in biology.	5	Origin: microbiology, OS security, organizational sociology.

[48] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[49] https://www.stratosphereips.org/blog/2026/2/4/rethinking-cybersecurity-immunity
[50] https://www.sentar.com/cybersecurity-building-resilience-in-the-digital-immune-system/
[51] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[52] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[53] https://www.stratosphereips.org/blog/2026/2/4/rethinking-cybersecurity-immunity
[54] https://www.sei.cmu.edu/blog/why-cybersecurity-is-not-like-the-immune-system/
[55] https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2016.00017/full
[56] https://pmc.ncbi.nlm.nih.gov/articles/PMC11458441/

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	"innate" immunity for wetware. Defenses located inside the system boundary.					
Level (of Immunity)	Tier in the proposed hierarchy of defensive mechanisms	Ranges from molecular recognition up through innate and adaptive immunity.	Parallel layers in information and social systems, from basic filters to collective responses.	Provides the document's scaffold: each level has analogous bio and info instances and maladaptations.	5	Origin: this work's multi-level Framework.
Localized	Bounded in physical or ideation space	Located at a specific subcellular compartment, membrane, or tissue region.	Opposite of distributed or decentralized, as in "Distributed Computing"	"localized" fundamentally indexes a constraint on information access		Can capture individualization or specialization of components or processes
Maladaptation	Once fit mechanism that now harms fitness.	Autoimmunity, allergy, chronic inflammation that reduce organismal fitness. [57] [58]	Tribalism, over-restrictive security, or brittle rules that damage group or system performance. [59] [60]	Every level's defense has characteristic failure modes; these are central to the analysis.	5	Origin: evolutionary biology, clinical immunology, social systems.
Marker / Signal	Observable indicator carrying identity or state information.	Surface markers, cytokines, and chemokines conveying immune information. [61]	Certificates, tokens, usage patterns; social signals like language and dress. [62] [63]	Emphasizes that identity and state are communicated via recognizable signals in all systems.	3	Origin: cell biology, signaling theory, communication theory.
Memory	Retained traces of past	Long-lived B/T memory cells enabling faster,	Threat intel, model parameters, and institutional or	Memory marks the transition from purely	4	Origin: immunology, ML,

[57] https://en.wikipedia.org/wiki/Immune_tolerance
[58] https://www.immunopaedia.org.za/immunology/special-focus-area/6-tolerance-and-autoimmunity/
[59] https://www.sei.cmu.edu/blog/why-cybersecurity-is-not-like-the-immune-system/
[60] https://www.stratosphereips.org/blog/2026/2/4/rethinking-cybersecurity-immunity
[61] https://ui.adsabs.harvard.edu/abs/2002Sci...296..298M/abstract
[62] https://oecs.mit.edu/pub/qlm9zp9e
[63] https://www.sciencedirect.com/topics/psychology/social-identity-theory

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	interactions that shape future responses.	stronger secondary responses. [64] [65]	cultural memory guiding later behavior. [66]	reactive to history-informed defense at higher levels.		organizational studies.
Mutation	Variation introduced into replicating entities.	Genetic and somatic mutations altering sequences and phenotypes. [67] [68]	Code changes, random perturbations in evolutionary algorithms, polymorphic malware. [69] [70]	Shared role as both source of innovation and mechanism of evasion.	3	Origin: genetics, software and algorithm design.
Negative Selection	Removing elements that overreact to self.	Deletion of self-reactive T and B cells during development. [71] [72]	AIS training that discards detectors matching "self" data; social exclusion of norm violators. [73] [74]	One half of a general calibration scheme (with positive selection) for self/nonself boundaries.	4	Origin: immunology, AIS, norm enforcement.
Pathogen	Entity that exploits a host and causes damage.	Disease-causing viruses, bacteria, fungi, and parasites. [75] [76]	Malware, hostile actors, or destabilizing narratives inside information and social systems. [77] [78]	General threat archetype whose replication and evasion strategies rhyme across substrates.	4	Origin: microbiology, cybersecurity, misinformation studies.
Plasmid	Small, transferable	Circular DNA carrying	Analogy for transferable code	Serves as a vivid example	2	Origin: microbiology, used

[64] https://www.immunopaedia.org.za/breaking-news/memory-and-immune-system-parallels-insights-into-adaptive-immunity-and-b-cell-responses/
[65] https://elifesciences.org/articles/26754
[66] https://pmc.ncbi.nlm.nih.gov/articles/PMC4070499/
[67] https://www.sciencedirect.com/topics/immunology-and-microbiology/horizontal-gene-transfer
[68] https://pmc.ncbi.nlm.nih.gov/articles/PMC3043529/
[69] https://pmc.ncbi.nlm.nih.gov/articles/PMC4070499/
[70] https://writings.stephenwolfram.com/2024/05/why-does-biological-evolution-work-a-minimal-model-for-biological-evolution-and-other-adaptive-processes/
[71] https://www.nature.com/articles/srep00769
[72] https://www.sciencedirect.com/science/article/pii/S107476130600358X
[73] https://www.cs.kent.ac.uk/pubs/2002/1343/content.pdf
[74] http://congres.cran.univ-lorraine.fr/2002/WCCI2002/CEC02/PDFFiles/Papers/8824.PDF
[75] https://my.clevelandclinic.org/health/articles/pathogen
[76] https://en.wikipedia.org/wiki/Virus
[77] https://pmc.ncbi.nlm.nih.gov/articles/PMC2732467/
[78] https://www.linkedin.com/pulse/biological-nature-cyber-attacks-unveiling-digital-pathogens-nair-jnqec

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	genetic element in microbes.	accessory genes, often spread via HGT. [79][80]	or content modules conferring new capabilities. [81][82]	of horizontal transfer across hosts and systems.		metaphorically in info systems.
Polarization	Splitting into opposed, reinforcing camps.	Cell/tissue polarity (different concept) in biology; limited immune usage.	Ideological or group sorting into mutually hostile camps under SGI. [83][84]	Treated as social-scale "autoimmunity," where group defense harms its own cognitive diversity.	4	Origin: political psychology, mapped onto immune metaphor.
Positive Selection	Keeping elements that can respond appropriatel y.	Retention of T cells that recognize self-MHC with moderate affinity. [85]	Selection of detectors that match important patterns; reward of identity-aligned behaviors. [86]	Complements negative selection: together they shape a functional recognition repertoire.	4	Origin: immunology, AIS, social reinforcement.
Receptor	Structure that binds to and senses specific inputs.	Immune receptors (TCR, BCR, TLR) sensing antigens or PAMPs. [87][88]	Conceptual receptors in AIS and sensors in distributed systems. [89][90]	Abstracts the interface layer where environment is sampled for recognition decisions.	3	Origin: immunology, sensor/ML design.
Recognition	Act of detecting and categorizing	Binding of receptors to antigens or PAMPs to	Pattern recognition in ML and security; social recognition of identity cues and group markers. [93][94]	Same functional step (pattern match → decision) across	5	Origin: immunology, pattern recognition, social cognition.

[79] https://pmc.ncbi.nlm.nih.gov/articles/PMC3043529/
[80] https://pmc.ncbi.nlm.nih.gov/articles/PMC22375/
[81] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[82] https://pmc.ncbi.nlm.nih.gov/articles/PMC11458441/
[83] https://www.simplypsychology.org/social-identity-theory.html
[84] https://www.britannica.com/topic/social-identity-theory
[85] https://www.nature.com/articles/srep00769
[86] https://www.cs.kent.ac.uk/pubs/2002/1343/content.pdf
[87] https://ui.adsabs.harvard.edu/abs/2002Sci...296..298M/abstract
[88] https://pmc.ncbi.nlm.nih.gov/articles/PMC61453/
[89] https://www.cs.kent.ac.uk/pubs/2002/1343/content.pdf
[90] https://www.scitepress.org/Papers/2025/132935/132935.pdf
[93] https://arxiv.org/html/2405.02325v4
[94] https://www.ebsco.com/research-starters/psychology/social-identity-theory

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	a pattern or entity.	trigger immune responses. [9192]		molecules, data, and social cues.		
Regulation / Suppression	Controls that limit and terminate responses.	Tregs, inhibitory cytokines, and apoptosis restraining immune responses. [9596]	Rate limits, overrides, norms and laws that prevent overreaction or cascade failures. [97 98]	Emphasizes that "turning off" defense is as important as detection and activation.	4	Origin: immunology, safety engineering, legal/institutional design.
Replication / Reproduction	Copying of entities over time.	Cell division, viral replication, and immune clonal expansion. [99100]	Copying of code, data, bots, and memes across networks. [101102]	Parallel replication logics enable both growth and spread of threats in different media.	3	Origin: virology, computing, memetics.
Response (Immune/ System)	Actions taken once a threat is recognized.	Inflammatory and adaptive cascades (antibodies, cytotoxic killing, complement). [103 104]	Incident response playbooks, lockouts, messaging, or mobilization in groups. [105106]	Cross-level concern with proportionality and timing of reactions, not just detection.	4	Origin: immunology, security operations, collective behavior.
Self	Operational boundary of what a system	Molecules and cells tolerated by the immune system as	Trusted users/processes in a system; psychological self	Defined by discrimination against nonself at molecular, organismal,	5	Origin: immunology, social psychology, security architecture.

[91] https://ui.adsabs.harvard.edu/abs/2002Sci...296..298M/abstract
[92] https://pmc.ncbi.nlm.nih.gov/articles/PMC61453/
[95] https://www.immunopaedia.org.za/immunology/special-focus-area/6-tolerance-and-autoimmunity/
[96] https://en.wikipedia.org/wiki/Immune_tolerance
[97] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[98] https://pmc.ncbi.nlm.nih.gov/articles/PMC11458441/
[99] https://en.wikipedia.org/wiki/Virus
[100] https://my.clevelandclinic.org/health/articles/pathogen
[101] https://pmc.ncbi.nlm.nih.gov/articles/PMC2732467/
[102] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[103] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[104] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[105] https://www.stratosphereips.org/blog/2026/2/4/rethinking-cybersecurity-immunity
[106] https://www.sentar.com/cybersecurity-building-resilience-in-the-digital-immune-system/

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	treats as "itself."	belonging to the organism. [107]	and group-based "social self." [108][109]	social, and digital levels.		
Signal / Noise	Valuable information vs. background.	Meaningful immune or physiological signals against molecular background. [110]	Useful data vs. random or irrelevant activity in security and communication. [111]	Highlights recognition's challenge: find true threats in a sea of normal variation.	2	Origin: signal processing, applied to immune and cyber contexts.
Social Group Identity (SGI)	Group-level mechanism defining "us" vs. "them."	Social organisms treating harm to group members as harm to self, driving coordinated defense.	Human sense of "we" that filters ideas and actors based on group membership rather than content. [112][113]	Treated as a social-scale immune system in "idea space," with analogues to cellular self/nonself.	5	Origin: social psychology (SIT/SCT), extended by this theory.
Specificity	Narrowness of what a recognition mechanism will match.	Highly specific binding in adaptive immunity to particular antigens. [114][115]	Classifier specificity (true negative rate) and precise matching in signatures.	Used to compare broad innate vs. narrow adaptive detection across systems.	3	Origin: immunology, statistics/ML.
Threshold	Activation point beyond which a response occurs.	Minimum stimulus needed to trigger immune activation; too low or high causes pathology. [116]	Detection cutoffs in anomaly detection, alert thresholds, social tipping points. [117][118]	Tradeoff between sensitivity and false alarms is structurally the same across domains.	3	Origin: immunology, ML, social dynamics.
Tolerance	Calibrated non-respons	Non-reactivity to self-antigens (central and	Social acceptance of difference; fault tolerance in	Links immunological "non-attack on	4	Origin: immunology, social

[107] https://pmc.ncbi.nlm.nih.gov/articles/PMC3136900/
[108] https://oecs.mit.edu/pub/qlm9zp9e
[109] https://www.simplypsychology.org/social-identity-theory.html
[110] https://ui.adsabs.harvard.edu/abs/2002Sci...296..298M/abstract
[111] https://pmc.ncbi.nlm.nih.gov/articles/PMC4384894/
[112] https://www.ebsco.com/research-starters/psychology/social-identity-theory
[113] https://www.simplypsychology.org/social-identity-theory.html
[114] https://www.ncbi.nlm.nih.gov/books/NBK27090/
[115] https://www.criver.com/eureka/immunology-non-immunologists-innate-vs-adaptive-immunity
[116] https://www.sciencedirect.com/science/article/pii/S107476130600358X
[117] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[118] https://pmc.ncbi.nlm.nih.gov/articles/PMC11458441/

Term	Definition	Biological usage	Informational/ social usage	Cross disciplinary note	Rating	Comment
	e to certain stimuli.	peripheral tolerance). [119][120]	systems continuing operation under failure.	self" with social and technical forbearance toward benign deviations.		theory, reliability engineering.
Vector	Carrier that delivers something into a new host or space.	Mosquitoes and other organisms transmitting pathogens; DNA vectors delivering genes. [121]	Attack vectors as paths into systems; numeric vectors in data models. [122][123]	Used to stress structural similarity of "delivery channels" in biology and cyber.	2	Origin: epidemiology, molecular biology, cybersecurity.
Virus	Minimal replicating agent that hijacks host machinery.	Viruses replicating inside host cells, often rapidly mutating. [124][125]	Self-replicating malicious code; "viral" content spreading in social media. [126][127]	Classical metaphor linking biological and digital contagion and spread.	3	Origin: virology, then adopted in computing and media theory.
Vulnerability	Exploitable weakness.	Immune deficiencies or genetic predispositions increasing disease risk. [128]	Software flaws, misconfigurations, and social fragilities exploitable by attackers. [129][130]	Conceptual complement of immunity at all scales—low immunity implies high vulnerability.	3	Origin: risk analysis across bio, cyber, and social domains.

[119] https://en.wikipedia.org/wiki/Immune_tolerance
[120] https://study.com/academy/lesson/immunologic-tolerance-definition-example.html
[121] https://my.clevelandclinic.org/health/articles/pathogen
[122] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[123] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[124] https://en.wikipedia.org/wiki/Virus
[125] https://my.clevelandclinic.org/health/articles/pathogen
[126] https://www.malwarebytes.com/blog/news/2017/10/our-computers-ourselves-digital-vs-biological-security
[127] https://www.linkedin.com/pulse/biological-nature-cyber-attacks-unveiling-digital-pathogens-nair-jnqec
[128] https://www.immunopaedia.org.za/immunology/special-focus-area/6-tolerance-and-autoimmunity/
[129] https://www.sciencedirect.com/science/article/am/pii/S0025556423000652
[130] https://pmc.ncbi.nlm.nih.gov/articles/PMC2732467/

References

Akerlof, G. A., & Kranton, R. E. (2000). Economics and identity. *The Quarterly Journal of Economics*, *115*(3), 715–753. https://doi.org/10.1162/003355300554881

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2022). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. In *arXiv [q-bio.NC]*. arXiv. https://doi.org/10.48550/arXiv.2212.14787

Alberts, B., Heald, R., Johnson, A., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2022). *Molecular biology of the cell*. W. W. Norton & Company.

Anderson, M. S., Venanzi, E. S., Klein, L., Chen, Z., Berzins, S. P., Turley, S. J., von Boehmer, H., Bronson, R., Dierich, A., Benoist, C., & Mathis, D. (2002). Projection of an immunological self shadow within the thymus by the aire protein. *Science (New York, N.Y.)*, *298*(5597), 1395–1401. https://doi.org/10.1126/science.1075958

Anderson, R. M., & May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature*, *318*(6044), 323–329. https://doi.org/10.1038/318323a0

Anil, C., Durmus, E., Rimsky, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D. J., Mosconi, F., Agrawal, R., Schaeffer, R., Bashkansky, N., Svenningsen, S., Lambert, M., Radhakrishnan, A., Denison, C., Hubinger, E. J., … Duvenaud, D. (2024). *Many-shot Jailbreaking*. https://openreview.net/forum?id=cw5mgd71jW

Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews. Neuroscience*, *10*(6), 410–422. https://doi.org/10.1038/nrn2648

Arumugam, T. V., Shiels, I. A., Woodruff, T. M., Granger, D. N., & Taylor, S. M. (2004). The role of the complement system in ischemia-reperfusion injury. *Shock (Augusta, Ga.)*, *21*(5), 401–409. https://doi.org/10.1097/00024382-200405000-00002

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, *11*(20), eadu9368. https://doi.org/10.1126/sciadv.adu9368

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). Constitutional AI: Harmlessness from AI Feedback. In *ArXiv*. Anthropic; arXiv. https://doi.org/10.48550/arXiv.2212.08073

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1607.06450

Baluška, F., & Levin, M. (2016). On having no head: Cognition throughout biological systems. *Frontiers in Psychology*, *7*, 902. https://doi.org/10.3389/fpsyg.2016.00902

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *10*(3), 295–307.

https://doi.org/10.1093/cercor/10.3.295
Bérut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., & Lutz, E. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, *483*(7388), 187–189. https://doi.org/10.1038/nature10872
Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *The Journal of Political Economy*, *100*(5), 992–1026. https://doi.org/10.1086/261849
Boch, R., Shearer, D. A., & Stone, B. C. (1962). Identification of isoamyl acetate as an active component in the sting pheromone of the honey bee. *Nature*, *195*(4845), 1018–1020. https://doi.org/10.1038/1951018b0
Bretscher, P., & Cohn, M. (1970). A theory of self-nonself discrimination. *Science*, *169*(3950), 1042–1049. https://doi.org/10.1126/science.169.3950.1042
Broz, P., & Dixit, V. M. (2016). Inflammasomes: mechanism of assembly, regulation and signalling. *Nature Reviews. Immunology*, *16*(7), 407–420. https://doi.org/10.1038/nri.2016.58
Bruneau, G. (n.d.). *McAfee DAT 5958 Update Issues*. McAfee. Retrieved February 16, 2026, from https://isc.sans.edu/diary/McAfee+DAT+5958+Update+Issues/8656/
Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., & Wu, J. (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2312.09390
Busso, N., & So, A. (2010). Mechanisms of inflammation in gout. *Arthritis Research & Therapy*, *12*(2), 206. https://doi.org/10.1186/ar2952
Calhoun, J. B. (1973). *From mice to men*. https://johnbcalhoun.com/wp-content/uploads/2019/01/1973-from-mice-to-men-secure.pdf
Cerf, V., & Kahn, R. (2021). A protocol for packet network intercommunication (1974). In *Ideas That Created the Future* (pp. 373–386). The MIT Press. https://doi.org/10.7551/mitpress/12274.003.0040
Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies: Controversies in Science & the Humanities*, *2*(3), 200–219. https://consc.net/papers/facing.html
Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. In *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/1712.05526
Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning models don't always say what they think. In *ArXiv*. arXiv. https://arxiv.org/abs/2505.05410
Chen, Y., & Malin, B. (2011). Detection of anomalous insiders in collaborative environments via relational analysis of access logs. *CODASPY: Proceedings of the ACM Conference on Data and Application Security and Privacy*, *2011*, 63–74. https://doi.org/10.1145/1943513.1943524
Cheswick, W. R., Bellovin, S. M., & Rubin, A. D. (2003). *Firewalls and Internet Security: Repelling the Wily Hacker*. Addison-Wesley.
Cikara, M., Bruneau, E., Van Bavel, J. J., & Saxe, R. (2014). Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of Experimental Social Psychology*, *55*, 110–125. https://doi.org/10.1016/j.jesp.2014.06.007
Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19.

https://doi.org/10.1093/analys/58.1.7
Cohn, M. (2010). The evolutionary context for a self-nonself discrimination. *Cellular and Molecular Life Sciences*, *67*(17), 2851–2862. https://doi.org/10.1007/s00018-010-0438-z
Common Sense Guide to Mitigating Insider Threats, Fifth Edition. (n.d.). Retrieved February 17, 2026, from https://www.odni.gov/files/NCSC/documents/nittf/20180209-CERT-Common-Sense-Guide-Fifth-Edition.pdf
Conrad, N., Misra, S., Verbakel, J. Y., Verbeke, G., Molenberghs, G., Taylor, P. N., Mason, J., Sattar, N., McMurray, J. J. V., McInnes, I. B., Khunti, K., & Cambridge, G. (2023). Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *Lancet*, *401*(10391), 1878–1890. https://doi.org/10.1016/S0140-6736(23)00457-9
Contos, B. (n.d.). *Environmental Drift Yields Cybersecurity Ineffectiveness*. ISACA. Retrieved February 16, 2026, from https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2019/environmental-drift-yields-cybersecurity-ineffectiveness
Cremer, S., Armitage, S. A. O., & Schmid-Hempel, P. (2007). Social immunity. *Current Biology*, *17*(16), R693–R702. https://doi.org/10.1016/j.cub.2007.06.008
CrowdStrike. (2024, September 21). *Technical Details: Falcon Update for Windows Hosts*. CrowdStrike.com. https://www.crowdstrike.com/en-us/blog/falcon-update-for-windows-hosts-technical-details/
Curio, E. (1988). Cultural transmission of enemy recognition by birds. In T. R. G. Zentall B (Ed.), *Social Learning: Psychological and Biological Perspectives* (pp. 75–97). Lawrence Erlbaum.
Cusick, M. F., Libbey, J. E., & Fujinami, R. S. (2012). Molecular mimicry as a mechanism of autoimmune disease. *Clinical Reviews in Allergy & Immunology*, *42*(1), 102–111. https://doi.org/10.1007/s12016-011-8294-7
Dalmau, J., & Rosenfeld, M. R. (2008). Paraneoplastic syndromes of the CNS. *Lancet Neurology*, *7*(4), 327–340. https://doi.org/10.1016/S1474-4422(08)70060-7
Damasio, A. (2000). *The feeling of what happens*. Vintage.
Damian, R. T. (1964). Molecular mimicry: Antigen sharing by parasite and host and its consequences. *The American Naturalist*, *98*(900), 129–149. https://doi.org/10.1086/282313
Davies, N. B., & Brooke, M. D. L. (1989). An Experimental Study of Co-Evolution Between the Cuckoo, Cuculus canorus, and its Hosts. I. Host Egg Discrimination. *The Journal of Animal Ecology*, *58*(1), 207. https://doi.org/10.2307/4995
Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446
Deneubourg, J.-L., Aron, S., Goss, S., & Pasteels, J. M. (1990). The self-organizing exploratory pattern of the argentine ant. *Journal of Insect Behavior*, *3*(2), 159–168. https://doi.org/10.1007/bf01417909
Derbinski, J., Schulte, A., Kyewski, B., & Klein, L. (2001). Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self. *Nature Immunology*, *2*(11),

1032–1039. https://doi.org/10.1038/ni723
Des Marais, D. J. (2003). Biogeochemistry of hypersaline microbial mats illustrates the dynamics of modern microbial ecosystems and the early evolution of the biosphere. *The Biological Bulletin*, *204*(2), 160–167. https://doi.org/10.2307/1543552
Detrain, C., & Deneubourg, J.-L. (2008). Collective decision-making and foraging patterns in ants and honeybees. In *Advances in Insect Physiology* (pp. 123–173). Elsevier. https://doi.org/10.1016/s0065-2806(08)00002-7
Dhabhar, F. S. (2014). Effects of stress on immune function: the good, the bad, and the beautiful. *Immunologic Research*, *58*(2-3), 193–210. https://doi.org/10.1007/s12026-014-8517-0
Ding, Z., Wang, W., Li, X., Wang, X., Jeon, G., Zhao, J., & Mu, C. (2024). Identifying alternately poisoning attacks in federated learning online using trajectory anomaly detection method. *Scientific Reports*, *14*(1), 20269. https://doi.org/10.1038/s41598-024-70375-w
Döring, Y., Libby, P., & Soehnlein, O. (2020). Neutrophil extracellular traps participate in cardiovascular diseases: Recent experimental and clinical insights: Recent experimental and clinical insights. *Circulation Research*, *126*(9), 1228–1241. https://doi.org/10.1161/CIRCRESAHA.120.315931
DuBois, E. F. (1937). *The Mechanism of Heat Loss and Temperature Regulation*. Stanford University Press.
Dunn, G. P., Old, L. J., & Schreiber, R. D. (2004). The three Es of cancer immunoediting. *Annual Review of Immunology*, *22*(1), 329–360. https://doi.org/10.1146/annurev.immunol.22.012703.104803
Dykstra, J., Gordon, L. A., Loeb, M. P., & Zhou, L. (2023). Maximizing the benefits from sharing cyber threat intelligence by government agencies and departments. *Journal of Cybersecurity*, *9*(1). https://doi.org/10.1093/cybsec/tyad003
Eckstein, T. (2023). The molecular heart of collective behavior in Dictyostelium. *Nature Communications*.
Elmore, S. (2007). Apoptosis: a review of programmed cell death. *Toxicologic Pathology*, *35*(4), 495–516. https://doi.org/10.1080/01926230701320337
Erlebacher, A. (2013). Immunology of the maternal-fetal interface. *Annual Review of Immunology*, *31*(1), 387–411. https://doi.org/10.1146/annurev-immunol-032712-100003
Fajgenbaum, D. C., & June, C. H. (2020). Cytokine storm. *The New England Journal of Medicine*, *383*(23), 2255–2273. https://doi.org/10.1056/NEJMra2026131
Fang, M., Cao, X., Jia, J., & Gong, N. (2020). Local Model Poisoning Attacks to {Byzantine-Robust} Federated Learning. *29th USENIX Security Symposium (USENIX Security 20)*, 1605–1622. https://www.usenix.org/conference/usenixsecurity20/presentation/fang
Flemming, H.-C., Wingender, J., Szewzyk, U., Steinberg, P., Rice, S. A., & Kjelleberg, S. (2016). Biofilms: an emergent form of bacterial life. *Nature Reviews. Microbiology*, *14*(9), 563–575. https://doi.org/10.1038/nrmicro.2016.94
Forrest, S., Allen, L., Perelson, A. S., & Cherukuri, R. (1994). Self-nonself discrimination in a computer. *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, 202–212. https://asu.elsevierpure.com/en/publications/self-nonself-discrimination-in-a-computer
Forrest, S., & Beauchemin, C. (2007). Computer immunology. *Immunological Reviews*, *216*(1), 176–197. https://doi.org/10.1111/j.1600-065X.2007.00499.x
Forrest, S., Hofmeyr, S. A., Somayaji, A., & Longstaff, T. A. (1996). A Sense of Self for Unix Processes. *Proceedings of the 1996 IEEE Symposium on Security and Privacy*, 120.

https://dl.acm.org/doi/10.5555/525080.884258
Forrest, S., Somayaji, A., & Ackley, D. (1997). Building Diverse Computer Systems. *Proceedings of the 6th Workshop on Hot Topics in Operating Systems (HotOS-VI)*, 67. https://dl.acm.org/doi/10.5555/822075.822408
Franceschi, C., Garagnani, P., Parini, P., Giuliani, C., & Santoro, A. (2018). Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nature Reviews. Endocrinology*, *14*(10), 576–590. https://doi.org/10.1038/s41574-018-0059-4
Franks, J., & Stolz, J. F. (2009). Flat laminated microbial mat communities. *Earth-Science Reviews*, *96*(3), 163–172. https://doi.org/10.1016/j.earscirev.2008.10.004
Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
Galli, S. J., & Tsai, M. (2012). IgE and mast cells in allergic disease. *Nature Medicine*, *18*(5), 693–704. https://doi.org/10.1038/nm.2755
Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. https://doi.org/10.1145/2523813
Gartner for Information Technology (IT) Leaders. (n.d.). Gartner. Retrieved February 16, 2026, from https://www.gartner.com/en/articles/what-is-a-digital-immune-system-and-why-does-it-matter
Godoy, L. D., Rossignoli, M. T., Delfino-Pereira, P., Garcia-Cairasco, N., & de Lima Umeoka, E. H. (2018). A comprehensive overview on stress neurobiology: Basic concepts and clinical implications. *Frontiers in Behavioral Neuroscience*, *12*, 127. https://doi.org/10.3389/fnbeh.2018.00127
Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. In *ArXiv*. arXiv. http://arxiv.org/abs/1701.00160
Goronzy, J. J., & Weyand, C. M. (2013). Understanding immunosenescence to improve responses to vaccines. *Nature Immunology*, *14*(5), 428–436. https://doi.org/10.1038/ni.2588
Gostic, K. M., Bridge, R., Brady, S., Viboud, C., Worobey, M., & Lloyd-Smith, J. O. (2019). Childhood immune imprinting to influenza A shapes birth year-specific risk during seasonal H1N1 and H3N2 epidemics. *PLoS Pathogens*, *15*(12), e1008109. https://doi.org/10.1371/journal.ppat.1008109
Goyal, A., Pal, O., Sundaram, H., Chandrasekharan, E., & Saha, K. (2026). Social simulacra in the wild: AI agent communities on Moltbook. In *ArXiv*. arXiv. https://doi.org/10.48550/arXiv.2603.16128
Graziano, M. (2015). *Consciousness and the social brain*. Oxford University Press.
Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). Alignment faking in large language models. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2412.14093
Griffin, A. S., & Galef, B. G., Jr. (2005). Social learning about predators: does timing matter? *Animal Behaviour*, *69*(3), 669–678. https://doi.org/10.1016/j.anbehav.2004.05.020
Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science (New York, N.Y.)*, *381*(6656), 398–404.

https://doi.org/10.1126/science.abp9364
Halstead, S. B. (2014). Dengue antibody-dependent enhancement: Knowns and unknowns. *Microbiology Spectrum*, *2*(6), AID – 0022–2014. https://doi.org/10.1128/microbiolspec.aid-0022-2014
Henter, J.-I., Horne, A., Aricó, M., Egeler, R. M., Filipovich, A. H., Imashuku, S., Ladisch, S., McClain, K., Webb, D., Winiarski, J., & Janka, G. (2007). HLH-2004: Diagnostic and therapeutic guidelines for hemophagocytic lymphohistiocytosis. *Pediatric Blood & Cancer*, *48*(2), 124–131. https://doi.org/10.1002/pbc.21039
Hoffman, H. M., Mueller, J. L., Broide, D. H., Wanderer, A. A., & Kolodner, R. D. (2001). Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. *Nature Genetics*, *29*(3), 301–305. https://doi.org/10.1038/ng756
Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an artificial immune system. *Evolutionary Computation*, *8*(4), 443–473. https://doi.org/10.1162/106365600568257
Hölldobler, B., & Wilson, E. O. (2009). *The Superorganism: The Beauty, Elegance, and Strangeness of Insect Societies*. W.W. Norton & Company.
Hooper, L. V., Littman, D. R., & Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science (New York, N.Y.)*, *336*(6086), 1268–1273. https://doi.org/10.1126/science.1223490
Howes, M. (2008). Self and nonself. In S. Sarkar & A. Plutynski (Eds.), *A Companion to the Philosophy of Biology* (pp. 271–286). Blackwell. https://www.lehigh.edu/~mhb0/PhilBioPotentialReadings%2020Oct18/Sahotra_Sarkar,_Anya_PlutynskiSelfNonself_A_Companion_to_th%20.pdf
Hubinger, E. (2024, December 1). Model Organisms of Misalignment. *AXRP - the AI X-Risk Research Podcast*. https://axrp.net/episode/2024/12/01/episode-39-evan-hubinger-model-organisms-misalignment.html
Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., … Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. In *ArXiv*. arXiv. http://arxiv.org/abs/2401.05566
Hu, Q., Yu, S.-Y., & Asghar, M. R. (2020). Analysing performance issues of open-source intrusion detection systems in high-speed networks. *Journal of Information Security and Applications*, *51*(102426), 102426. https://doi.org/10.1016/j.jisa.2019.102426
Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*(1), 653–670. https://doi.org/10.1146/annurev.psych.60.110707.163604
Iba, T., Watanabe, E., Umemura, Y., Wada, T., Hayashida, K., Kushimoto, S., Japanese Surviving Sepsis Campaign Guideline Working Group for disseminated intravascular coagulation, & Wada, H. (2019). Sepsis-associated disseminated intravascular coagulation and its differential diagnoses. *Journal of Intensive Care*, *7*(1), 32. https://doi.org/10.1186/s40560-019-0387-z
Internet Filters. (2016, August 5). National Coalition Against Censorship. https://ncac.org/resource/internet-filters-2
Janeway, C. A. (1999). *Immunobiology: The immune system in health and disease* (4th ed.). Garland Publishing.
Janeway, C. A., Jr, & Medzhitov, R. (2002). Innate immune recognition. *Annual Review of Immunology*, *20*(1), 197–216.

https://doi.org/10.1146/annurev.immunol.20.083001.084359
Janis, I. L. (1972). Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes. *Viii*, *277*. https://psycnet.apa.org/record/1975-29417-000
Jansen, W., & Grance, T. (2011). *Guidelines on security and privacy in public cloud computing*. National Institute of Standards and Technology. https://doi.org/10.6028/nist.sp.800-144
Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate Feedback Loops in Recommender Systems. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 383–390. https://doi.org/10.1145/3306618.3314288
Jiang, Y., Zhang, Y., Shen, X., Backes, M., & Zhang, Y. (2026). "Humans welcome to observe": A First Look at the Agent Social Network Moltbook. In *ArXiv*. arXiv. https://doi.org/10.48550/arXiv.2602.10127
Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., & Glaese, A. (2025). Training LLMs for Honesty via Confessions. In *arXiv [cs.LG]*. arXiv. https://doi.org/10.48550/arXiv.2512.08093
Johnson, N. L. (1998). *Collective Problem Solving: Functionality beyond the Individual* (Nos. LAUR-98-2227). Los Alamos National Laboratory. https://collectivescience.com/publications/
Johnson, N. L. (1999). Diversity in Decentralized Systems: Enabling Self-Organizing Solutions. In *Conference Proceedings of Decentralization Two*.
Johnson, N. L. (2023). Observations on modeling social identity: Suggestions to address the challenges of social identity. In *Advances in Social Simulation* (pp. 285–298). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-34920-1_23
Johnson, N. L. (2026a). *A Functional Theory of Ethical Behavior: ethics as adaptive self-other regulation, realized in four multi-level hypotheses, applied to AI ethical development*. https://doi.org/10.13140/RG.2.2.11124.51849
Johnson, N. L. (2026b). *Evolutionary Origins and Neurochemistry of Fight-or-Flight and Social Copying: Parallel Innate Survival Systems at Individual and Collective Levels*. http://collectivescience.com/social-identity
Johnson, N. L. (2026c). *Primer on Social Group Identity (SGI): The missing link in understanding human behavior, influence, and conflict (v4.4)*. https://collectivescience.com/social-identity/
Johnson, N. L. (2026d). *The biochemistry of collective survival and its consequences in modern culture*. https://collectivescience.com/wp-content/uploads/2026/02/NLJ-Biochemistry-of-Collective-Survival-and-Its-Consequences-in-Modern-Culture.pdf
Johnson, N. L. (2026e). *The Moltbook Singularity and the Evolution of Digital Immunity: (a rapid-release summary)*. https://collectivescience.com/social-identity/Moltbook
Kauffman, S. A. (1971). Cellular homeostasis, epigenesis and replication in randomly aggregated macromolecular systems. *Journal of Cybernetics*, *1*(1), 71–96. https://doi.org/10.1080/01969727108545830
Kim, J., Park, M., Kim, H., Cho, S., & Kang, P. (2019). Insider threat detection based on user behavior modeling and anomaly detection algorithms. *Applied Sciences (Basel, Switzerland)*, *9*(19), 4018. https://doi.org/10.3390/app9194018
Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kish, T. (2024, February 20). *SIEM Migration: Challenges and Strategies*. CardinalOps. https://cardinalops.com/blog/siem-migration-challenges-strategies/
Klein, L., Kyewski, B., Allen, P. M., & Hogquist, K. A. (2014). Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nature Reviews. Immunology*, *14*(6), 377–391. https://doi.org/10.1038/nri3667
Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews. Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22
Kurosaki, T., Kometani, K., & Ise, W. (2015). Memory B cells. *Nature Reviews. Immunology*, *15*(3), 149–159. https://doi.org/10.1038/nri3802
Kutasov, J., Sun, Y., Colognese, P., van der Weij, T., Petrini, L., Zhang, C. B. C., Hughes, J., Deng, X., Sleight, H., Tracy, T., Shlegeris, B., & Benton, J. (2025). SHADE-Arena: Evaluating sabotage and monitoring in LLM agents. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2506.15740
Labrie, S. J., Samson, J. E., & Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature Reviews. Microbiology*, *8*(5), 317–327. https://doi.org/10.1038/nrmicro2315
Lakkis, F. G., & Lechler, R. I. (2013). Origin and biology of the allogeneic response. *Cold Spring Harbor Perspectives in Medicine*, *3*(8), a014993–a014993. https://doi.org/10.1101/cshperspect.a014993
Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001
Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183–191.
Larsen, P., Homescu, A., Brunthaler, S., & Franz, M. (2014). SoK: Automated Software Diversity. *2014 IEEE Symposium on Security and Privacy*, 276–291. https://doi.org/10.1109/sp.2014.25
Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009
Ledford, H. (2022). Neurons in a dish learn to play Pong - what's next? *Nature*, *610*(7932), 433. https://doi.org/10.1038/d41586-022-03229-y
Lee, Y., Gong, J., & Kang, J. (2025). Embedding Byzantine fault tolerance into federated learning via consistency scoring. In *ArXiv*. arXiv. http://arxiv.org/abs/2411.10212
Levi, M., & van der Poll, T. (2017). Coagulation and sepsis. *Thrombosis Research*, *149*, 38–44. https://doi.org/10.1016/j.thromres.2016.11.007
Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-Free Cognition. *Frontiers in Psychology*, *10*, 2688. https://doi.org/10.3389/fpsyg.2019.02688
Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638
Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., & Zhang, X. (2018). Trojaning Attack on Neural Networks. *Proceedings 2018 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA. https://doi.org/10.14722/ndss.2018.23291
Liu, Y.-P., Wang, Y.-S., Zhan, B., Wang, R., & Jiang, Y. (2025). The influence of social context on perceptual decision making and its computational neural mechanisms. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Jin Zhan*, *52*(10), 2568–2584.

https://doi.org/10.3724/j.pibb.2025.0217
Li, Y., Tang, T., Hsieh, C.-J., & Lee, T. C. M. (2021). Adversarial examples detection with Bayesian neural network. In *ArXiv*. arXiv. http://arxiv.org/abs/2105.08620
Llewelyn, M., & Cohen, J. (2002). Superantigens: microbial agents that corrupt immunity. *The Lancet Infectious Diseases*, *2*(3), 156–162. https://doi.org/10.1016/s1473-3099(02)00222-0
Lodish, H., Berk, A., Kaiser, C., Krieger, M., Bretscher, A., Ploegh, H., Martin, K., Yaffe, M., & Amon, A. (2021). *Molecular cell biology* (9th ed.). W.H. Freeman.
Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 1–1. https://doi.org/10.1109/tkde.2018.2876857
MacDiarmid, M., Hubinger, E., & Anthropic Alignment Team. (2025). *Natural emergent misalignment from reward hacking in production RL*. Anthropic PBC. https://arxiv.org/abs/2511.18397
MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., Bouscal, N., Bowman, S., Bricken, T., Cloud, A., Denison, C., Gasteiger, J., Greenblatt, R., Leike, J., Lindsey, J., Mikulik, V., Perez, E., Rodrigues, A., Thomas, D., … Hubinger, E. (2025). Natural emergent misalignment from reward hacking in production RL. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2511.18397
Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's Law. In *ArXiv*. arXiv. http://arxiv.org/abs/1803.04585
Marée, A. F., & Hogeweg, P. (2001). How amoeboids self-organize into a fruiting body: multicellular coordination in Dictyostelium discoideum. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(7), 3879–3883. https://doi.org/10.1073/pnas.061535198
Martinon, F., Pétrilli, V., Mayor, A., Tardivel, A., & Tschopp, J. (2006). Gout-associated uric acid crystals activate the NALP3 inflammasome. *Nature*, *440*(7081), 237–241. https://doi.org/10.1038/nature04516
Martin, W., & Russell, M. J. (2007). On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1486), 1887–1925. https://doi.org/10.1098/rstb.2006.1881
Ma, S., Liu, Y., Tao, G., Lee, W.-C., & Zhang, X. (2019). NIC: Detecting adversarial samples with neural network invariant checking. *Proceedings 2019 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA. https://doi.org/10.14722/ndss.2019.23415
Mason, M. F., Dyer, R., & Norton, M. I. (2009). Neural mechanisms of social influence. *Organizational Behavior and Human Decision Processes*, *110*(2), 152–159. https://doi.org/10.1016/j.obhdp.2009.04.001
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company. https://doi.org/10.1007/978-94-009-8947-4
Matzinger, P. (1994). Tolerance, danger, and the extended family. *Annual Review of Immunology*, *12*(1), 991–1045. https://doi.org/10.1146/annurev.iy.12.040194.005015
Matzinger, P. (2002). The danger model: a renewed sense of self. *Science (New York, N.Y.)*, *296*(5566), 301–305. https://doi.org/10.1126/science.1071059
McLane, L. M., Abdel-Hakeem, M. S., & Wherry, E. J. (2019). CD8 T cell exhaustion during chronic viral infection and cancer. *Annual Review of Immunology*, *37*(1), 457–495. https://doi.org/10.1146/annurev-immunol-041015-055318
McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (20--22 Apr 2017).

Communication-Efficient Learning of Deep Networks from Decentralized Data. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

Merle, N. S., Church, S. E., Fremeaux-Bacchi, V., & Roumenina, L. T. (2015). Complement system part I â€“ molecular mechanisms of activation and regulation. *Frontiers in Immunology*, *6*, 262. https://doi.org/10.3389/fimmu.2015.00262

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, M. B., & Bassler, B. L. (2001). Quorum sensing in bacteria. *Annual Review of Microbiology*, *55*(1), 165–199. https://doi.org/10.1146/annurev.micro.55.1.165

Min, B. H., & Borch, C. (2022). Systemic failures and organizational risk management in algorithmic trading: Normal accidents and high reliability in financial markets. *Social Studies of Science*, *52*(2), 277–302. https://doi.org/10.1177/03063127211048515

MiniMax. (2026, May 31). *MiniMax M3: Frontier coding, 1M context, native multimodality — all in one model*. MiniMax. https://www.minimax.io/blog/minimax-m3

Mitnick, K. D., & Simon, W. L. (2001). *The art of deception: Controlling the human element of security*. Wiley.

Müller, M., Wandel, S., Colebunders, R., Attia, S., Furrer, H., Egger, M., & IeDEA Southern and Central Africa. (2010). Immune reconstitution inflammatory syndrome in patients starting antiretroviral therapy for HIV infection: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *10*(4), 251–261. https://doi.org/10.1016/S1473-3099(10)70026-8

Nilsson, A., Wilhelms, D. B., Mirrasekhian, E., Jaarola, M., Blomqvist, A., & Engblom, D. (2017). Inflammation-induced anorexia and fever are elicited by distinct prostaglandin dependent mechanisms, whereas conditioned taste aversion is prostaglandin independent. *Brain, Behavior, and Immunity*, *61*, 236–243. https://doi.org/10.1016/j.bbi.2016.12.007

Norman, A., Hansen, L. H., & Sørensen, S. J. (2009). Conjugative plasmids: vessels of the communal gene pool. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1527), 2275–2289. https://doi.org/10.1098/rstb.2009.0037

Orca security. (2022). *2022 Cloud Security Alert Fatigue Report*. Orca Security. https://orca.security/resources/blog/2022-cloud-security-alert-fatigue-report/

Ostrom, E. (2015). *Governing the Commons: The Evolution of Institutions for Collective Action (Canto Classics ed)*. Cambridge University Press.

Papayannopoulos, V. (2018). Neutrophil extracellular traps in immunity and disease. *Nature Reviews. Immunology*, *18*(2), 134–147. https://doi.org/10.1038/nri.2017.105

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). On the difficulty of training Recurrent Neural Networks. In *arXiv*. arXiv. http://arxiv.org/abs/1211.5063

Peng, Q., Li, K., Smyth, L. A., Xing, G., Wang, N., Meader, L., Lu, B., Sacks, S. H., & Zhou, W. (2012). C3a and C5a promote renal ischemia-reperfusion injury. *Journal of the American Society of Nephrology*, *23*(9), 1474–1485. https://doi.org/10.1681/ASN.2011111072

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., & Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science (New York, N.Y.)*, *271*(5255), 1582–1586. https://doi.org/10.1126/science.271.5255.1582

Pradeu, T., & Vivier, E. (2016). The discontinuity theory of immunity. *Science Immunology*, *1*(1). https://doi.org/10.1126/sciimmunol.aag0479

Price, H. C. W., AlMuhanna, H., Bassani, P. M., Ho, M., & Evans, T. S. (2026). Let there be claws: An early social network analysis of AI agents on Moltbook. In *arXiv [physics.soc-ph]*.

arXiv. https://doi.org/10.48550/arXiv.2602.20044
Proksch, E., Brandner, J. M., & Jensen, J.-M. (2008). The skin: an indispensable barrier. *Experimental Dermatology*, *17*(12), 1063–1072. https://doi.org/10.1111/j.1600-0625.2008.00786.x
Pross, A. (2012). *What is life?: How chemistry becomes biology*. Oxford University Press.
Qi, J., Li, M., Liu, J., Shu, Y., Yu, D., Ma, S., Cui, W., Zhao, Y., Chen, Y., Jiang, R., King, I., & Xu, Z. (2026). Towards trustworthy agentic AI: a comprehensive survey of safety, robustness, privacy, and system security. *Academia AI and Applications*, *2*(2). https://doi.org/10.20935/acadai8260
Riding, R. (2011). The nature of stromatolites: 3,500 million years of history and a century of research. In *Advances in Stromatolite Geobiology* (pp. 29–74). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10415-2_3
Rosenstein, M. T., Marx, Z., Kaelbling, L. P., & Dietterich, T. G. (2005). *To transfer or not to transfer*. Workshop on Inductive Transfer. https://web.engr.oregonstate.edu/~tgd/publications/rosenstein-marx-kaelbling-dietterich-hnb-nips2005-transfer-workshop.pdf
Rosenthal, D. (2005). *Consciousness and Mind*. Clarendon Press.
Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture*. https://doi.org/10.6028/nist.sp.800-207-draft2
Round, J. L., & Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology*, *9*(5), 313–323. https://doi.org/10.1038/nri2515
Ruiz-Mirazo, K., Briones, C., & de la Escosura, A. (2014). Prebiotic systems chemistry: new perspectives for the origins of life. *Chemical Reviews*, *114*(1), 285–366. https://doi.org/10.1021/cr2004844
Saeli, S., Bisio, F., Lombardo, P., & Massa, D. (2020). DNS covert channel detection via behavioral analysis: A machine learning approach. In *ArXiv*. arXiv. http://arxiv.org/abs/2010.01582
Sallusto, F., Geginat, J., & Lanzavecchia, A. (2004). Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annual Review of Immunology*, *22*(1), 745–763. https://doi.org/10.1146/annurev.immunol.22.012703.104702
Sapolsky, R. M. (2017). *Behave*. Penguin Press.
Schatz, D. G., & Swanson, P. C. (2011). V(D)J recombination: mechanisms of initiation. *Annual Review of Genetics*, *45*(1), 167–202. https://doi.org/10.1146/annurev-genet-110410-132552
Schevon, C. A., Weiss, S. A., McKhann, G., Jr, Goodman, R. R., Yuste, R., Emerson, R. G., & Trevelyan, A. J. (2012). Evidence of an inhibitory restraint of seizure activity in humans. *Nature Communications*, *3*(1), 1060. https://doi.org/10.1038/ncomms2056
Schreiber, R. D., Old, L. J., & Smyth, M. J. (2011). Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science (New York, N.Y.)*, *331*(6024), 1565–1570. https://doi.org/10.1126/science.1203486
Securities, U. S., Street, 1155 21st, Street, 100 F., Washington, N. E., C., D., & C., D. (n.d.). *Report of the staff of the cftc and sec to the joint advisory committee on emerging regulatory issues*. Retrieved February 17, 2026, from https://www.sec.gov/news/studies/2010/marketevents-report.pdf
Seth, A. (2021). *Being you*. Dutton.
Sharma, B., Pokharel, P., & Joshi, B. (2020, July). User behavior analytics for anomaly detection using LSTM autoencoder - insider threat detection. *Proceedings of the 11th International Conference on Advances in Information Technology*. IAIT2020: The 11th International

Conference on Advances in Information Technology, Bangkok Thailand. https://doi.org/10.1145/3406601.3406610
Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., Agarwal, R., Anil, C., Askell, A., Bailey, N., Benton, J., Bluemke, E., Bowman, S. R., Christiansen, E., Cunningham, H., Dau, A., Gopal, A., … Perez, E. (2025). Constitutional Classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2501.18837
Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., & Angus, D. C. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *The Journal of the American Medical Association*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287
Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward gaming. *Neural Information Processing Systems*, *35*, 9460–9471. https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html
Spribille, T., Resl, P., Stanton, D. E., & Tagirdzhanova, G. (2022). Evolutionary biology of lichen symbioses. *The New Phytologist*, *234*(5), 1566–1582. https://doi.org/10.1111/nph.18048
Sriram, K., Montgomery, D., McPherson, D. R., Osterweil, E., & Dickson, B. (2016). *RFC 7908: Problem Definition and Classification of BGP Route Leaks*. IETF Datatracker. https://datatracker.ietf.org/doc/html/rfc7908
Stallen, M., & Sanfey, A. G. (2015). The neuroscience of social conformity: implications for fundamental and applied research. *Frontiers in Neuroscience*, *9*, 337. https://doi.org/10.3389/fnins.2015.00337
Sun, R., Zhu, Y., Fei, J., & Chen, X. (2023). A survey on moving target defense: Intelligently affordable, optimized and self-adaptive. *Applied Sciences (Basel, Switzerland)*, *13*(9), 5367. https://doi.org/10.3390/app13095367
Sunstein, C. R. (2009). *Republic.Com 2.0*. Princeton University Press.
Szostak, J. W., Bartel, D. P., & Luisi, P. L. (2001). Synthesizing life. *Nature*, *409*(6818), 387–390. https://doi.org/10.1038/35053176
Tajfel, H., & Turner, J. (2000). An integrative theory of intergroup conflict. In *Organizational Identity* (pp. 56–65). Oxford University PressOxford. https://doi.org/10.1093/oso/9780199269464.003.0005
Takaba, H., Morishita, Y., Tomofuji, Y., Danks, L., Nitta, T., Komatsu, N., Kodama, T., & Takayanagi, H. (2015). Fezf2 orchestrates a thymic program of self-antigen expression for immune tolerance. *Cell*, *163*(4), 975–987. https://doi.org/10.1016/j.cell.2015.10.013
Tariq, S., Baruwal Chhetri, M., Nepal, S., & Paris, C. (2025). Alert fatigue in security operations centres: Research challenges and opportunities. *ACM Computing Surveys*, *57*(9), 1–38. https://doi.org/10.1145/3723158
Tauber, A. I. (2015). Reconceiving autoimmunity: An overview. *Journal of Theoretical Biology*, *375*, 52–60. https://doi.org/10.1016/j.jtbi.2014.05.029
Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Daniel Freeman, C., … Henighan, T. (2026). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2605.29358
Thanh-Tung, H., & Tran, T. (2020, July). Catastrophic forgetting and mode collapse in GANs.

2020 International Joint Conference on Neural Networks (IJCNN). 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom. https://doi.org/10.1109/ijcnn48605.2020.9207181

Three Sketches of ASL-4 Safety Case Components. (n.d.). Retrieved February 16, 2026, from https://alignment.anthropic.com/2024/safety-cases/

Toelch, U., & Dolan, R. J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. *Trends in Cognitive Sciences*, *19*(10), 579–589. https://doi.org/10.1016/j.tics.2015.07.007

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, *302*(5909), 575–581. https://doi.org/10.1038/302575a0

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews. Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behaviour*, *3*(2), 183–193. https://doi.org/10.1038/s41562-018-0518-x

Trevelyan, A. J., Sussillo, D., Watson, B. O., & Yuste, R. (2006). Modular propagation of epileptiform activity: evidence for an inhibitory veto in neocortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *26*(48), 12447–12455. https://doi.org/10.1523/JNEUROSCI.2787-06.2006

van der Poll, T., van de Veerdonk, F. L., Scicluna, B. P., & Netea, M. G. (2017). The immunopathology of sepsis and potential therapeutic targets. *Nature Reviews. Immunology*, *17*(7), 407–420. https://doi.org/10.1038/nri.2017.36

Van Valen, L. (1973). A New Evolutionary Law. *Evolutionary Theory*, *1*, 1–30. https://ebme.marine.rutgers.edu/HistoryEarthSystems/HistEarthSystems_Fall2010/VanValen%201973%20Evol%20%20Theor%20.pdf

Vatti, A., Monsalve, D. M., Pacheco, Y., Chang, C., Anaya, J.-M., & Gershwin, M. E. (2017). Original antigenic sin: A comprehensive review. *Journal of Autoimmunity*, *83*, 12–21. https://doi.org/10.1016/j.jaut.2017.04.008

Victora, G. D., & Nussenzweig, M. C. (2012). Germinal centers. *Annual Review of Immunology*, *30*(1), 429–457. https://doi.org/10.1146/annurev-immunol-020711-075032

Vinitsky, E., Köster, R., Agapiou, J. P., Duéñez-Guzmán, E. A., Vezhnevets, A. S., & Leibo, J. Z. (2023). A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, *2*(2), 263391372311620. https://doi.org/10.1177/26339137231162025

Vivier, E., Tomasello, E., Baratin, M., Walzer, T., & Ugolini, S. (2008). Functions of natural killer cells. *Nature Immunology*, *9*(5), 503–510. https://doi.org/10.1038/ni1582

Wang, Z., Dai, Z., Poczos, B., & Carbonell, J. (2019, June). Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. https://doi.org/10.1109/cvpr.2019.01155

Wikipedia contributors. (2026, February 10). *2024 CrowdStrike-related IT outages*. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/2024_CrowdStrike-related_IT_outages

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly Review of Biology*, *82*(4), 327–348. https://doi.org/10.1086/522809

Xavier, J. C., Hordijk, W., Kauffman, S., Steel, M., & Martin, W. F. (2020). Autocatalytic chemical networks at the origin of metabolism. *Proceedings. Biological Sciences*,

287(1922), 20192377. https://doi.org/10.1098/rspb.2019.2377
Xu, H., Yang, J., Gao, W., Li, L., Li, P., Zhang, L., Gong, Y.-N., Peng, X., Xi, J. J., Chen, S., Wang, F., & Shao, F. (2014). Innate immune sensing of bacterial modifications of Rho GTPases by the Pyrin inflammasome. *Nature*, *513*(7517), 237–241. https://doi.org/10.1038/nature13449
Yee, B., & Koh, P. (2026). Benchmarking emergent coordination in large-scale LLM populations: An evaluation framework on the MoltBook archive. In *ArXiv*. arXiv. https://doi.org/10.48550/arXiv.2603.03555
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416
Yuan, F., Cao, Y., Shang, Y., Liu, Y., Tan, J., & Fang, B. (2018). Insider threat detection with deep neural network. In *Lecture Notes in Computer Science* (pp. 43–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-93698-7_4
Zeto, J. (2021, July 26). *What is Deep Packet Inspection (DPI)?* Apposite Technologies. https://apposite-tech.com/what-is-deep-packet-inspection-dpi/
Zhang, Y., Mei, K., Liu, M., Wang, J., Metaxas, D. N., Wang, X., Hamm, J., & Ge, Y. (2026). Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook. In *ArXiv*. arXiv. https://doi.org/10.48550/arXiv.2602.13284