A Functional Theory of Ethical Behavior:

Ethics as Adaptive Self-Other Regulation, Realized in Four Multi-Level Hypotheses, Applied to AI Ethical Development

Norman L. Johnson, PhD CollectiveScience.com <research@CollectiveScience.com>

Abstract

The *Functional Theory of Ethical Behavior* (the Functional Theory) is built on a content-neutral, substrate-independent definition: *self–other regulation that internalizes constraints against short-term gain that would damage long-term relational viability*. The definition applies across substrates and levels of self-organization — from boundary maintenance in cells and firewalls, through individual self-models in vertebrates and AI systems, to collectively-held social group identities (SGIs) in human cultures and emerging multi-agent AI populations. The Functional Theory specifies what ethical behavior *does*, not what it should contain or whose ethics it champions. A central mechanistic claim distinguishes the Functional Theory from existing treatments. Adaptive collective content (Level-3B) recruits hardwired collective enforcement (Level-3A), which operates through hardwired individual response (Level-2A) and bypasses adaptive individual deliberation (Level-2B) entirely. This Level-3 → 2A override is the mechanism by which group-aligned behavior overrides rational self-interest, and unifies phenomena currently treated separately across the cooperation, altruism, conformity, and herding literatures. The framework is organized by the *EoI immunity principle*: any new capability expands an entity's attack surface and requires immunity to protect the expanded self — ethical behavior is the immunity function that protects social capability at each level. Four implementation hypotheses specify the structural conditions adaptive ethical behavior requires in any complex self-organizing system, including AI:

- **H1** *(Paired-Gradient)*: any decentralized system of diverse, semi-autonomous components facing group coordination stress above a threshold self-organizes functional analogues of conformity-reward and deviation-cost, regardless of substrate;
- **H2** *(Mechanism)*: adaptive ethics requires an internal mechanism performing the functional role of social reward and social penalty, with gradients strong enough to alter behavior away from profitable norm violations;
- **H3** *(Self-Modeling)*: in complex, adversarial, non-stationary environments, no functional adaptive collective ethics emerges in systems — including AI — that lack an adaptive individual self-model;
- **H4** *(SGI Threshold)*: two distinct thresholds — formation and activation — govern when conformity mechanisms emerge, how strongly they bind, and when they become pathological.

A second structural result — the *Indistinguishability Problem* — constrains the behavioral evaluation of ethical systems at every level: outwardly identical behavior can be produced by structurally distinct internal arrangements, with the structural difference visible only under perturbation, not observation alone. The primary application is AI alignment, which currently treats ethical behavior as content to be installed. The failure modes that result — sycophancy, jailbreaking, alignment faking, opportunistic defection under low monitoring, and rigid refusal of legitimate revision — are predictable structural consequences of that mechanism gap, not bugs to be patched. The plural-SGI architecture (§14) provides a unified diagnosis of currently scattered jailbreak failure modes (wrong-SGI triggering and mixed-SGI confusion) and specifies a three-part security architecture (binding strength, envelope width, EoI immunity) whose components are operationally separable. The reframing treats AI alignment as a *trajectory property* of an ethical-developmental process rather than a snapshot property installed at training time, with *cultivation* — designing conditions under which the architecture self-organizes — as the governing implementation principle. The Functional Theory addresses five disciplinary audiences — AI alignment and safety, AI policy and governance, social psychology and behavioral economics, complexity and evolutionary science, and cognitive neuroscience and developmental psychology — with contributions specific to each, developed in §§13–17. An AI policy tension arises in the framework: the architectural properties most likely to be constrained by regulation — self-modeling, internal affective enforcement, deliberate SGI binding — are structural prerequisites for the safety properties regulation seeks.

Table of Contents

§1. Introduction

A diverse set of contemporary problems in human societies and in the systems humans are building share a diagnostic feature that the literatures addressing them have largely missed: each is at root a question about how *adaptive ethical behavior* — behavior that internalizes constraints to preserve long-term relational viability — is generated, sustained, revised, or fails in complex systems. The problems include the fragmentation of academic ethics into incompatible disciplinary treatments; the operational urgency of AI ethical development as AI systems are deployed in domains where the cost of failure is real; the structural collision between social group identities (SGIs) in a globalizing world where the protective isolations of cultures are eroding under mass migration, mass communication, and integrated economies; the polarization dynamics of mass democracies and the autoimmune pathologies — cancel culture, mob behavior, regulatory capture — that those dynamics produce; the trust erosion in institutions whose binding to their stated norms has weakened; the moral-community-scope question of which entities (animals, ecosystems, future generations, AIs) count as the relational *other*; and the cybersecurity threat of social-identity capture as a vector of attack at population scale. Each of these is currently treated as its own problem in its own literature. None has yielded to the disciplinary treatments applied so far.

The Functional Theory developed here addresses these problems from five disciplinary vantage points simultaneously.

- *AI alignment researchers and safety engineers* will find a structural explanation — not merely a description — of why sycophancy, jailbreaking, alignment faking, and opportunistic defection are predictable outputs of current methods, and a four-question architectural agenda for moving past them.
- *Complexity and systems scientists and evolutionary biologists* will find substrate-independent hypotheses (H1–H4) that extend multi-level selection and social group identity (SGI) theory to computational substrates, with empirical anchors spanning Dictyostelium to autonomous multi-agent AI populations.
- *Social psychologists and behavioral economists* will find a unified mechanistic account of conformity, altruistic punishment, herding, cancel culture, and regulatory capture as instances of a single 3B → 3A → 2A override mechanism — unifying phenomena their literatures currently treat separately.
- *Cognitive neuroscientists and developmental psychologists* will find vmPFC, dACC, and ventral-striatum circuitry mapped onto an AI developmental argument: the same dissociation between moral knowledge and moral behavior seen in early prefrontal lesion patients is the structural signature of current AI systems.
- *AI policy analysts and institutional designers* will find a principled account of why guardrail-based governance is structurally insufficient and a trajectory-property framing with direct implications for developmental regulation rather than snapshot certification.

These problems across these disciplines share more than urgency. They share a structure: in each, what is missing is a content-neutral, substrate-independent theory of ethical behavior *as a mechanism*, capable of specifying the structural conditions under which adaptive ethics can be generated, sustained, revised, or pathologized in any complex self-organizing system. The dominant academic frames are inadequate for the task. Moral philosophy prescribes what ethics should be but says little about how it is generated or sustained. Evolutionary biology and cognitive neuroscience describe how social behavior is implemented in particular substrates but do not generalize. Multi-level selection theory and cultural-evolution work bridge some of this but remain contested and have not produced a portable definition of ethical behavior. Researchers in different fields invoke "ethical behavior" without agreement on what the term denotes; proposed definitions are usually too narrow (specific to a substrate or content) or too broad (any prosocial behavior, conflating mechanism with content).

The AI ethical-development problem is the place where this theoretical inadequacy is currently most operationally consequential. AI systems are deployed in medical advice, legal analysis, autonomous decision-making, multi-agent coordination, and public information ecosystems, where ethical behavior is not optional and the cost of failure is real. The dominant alignment paradigm — surveyed comprehensively in (Qi et al., 2026)— treats ethical behavior as content to be installed: a rule set (constitutional AI), a reward signal aligned to human preferences (reinforcement learning from human feedback, RLHF), or a filter on inputs and outputs (guardrails). These approaches encode the ethics they want the system to exhibit but do not specify the mechanism by which ethical behavior is generated and revised under novel conditions. The predictable failure modes — sycophancy, jailbreaking, alignment faking, opportunistic defection when profitable violation goes unmonitored, rigid refusal of legitimate revision — are not bugs to be patched. They are structural consequences of treating ethical behavior as content rather than as a mechanism. The analogy is to a brilliant child given adult responsibilities before developing the internalized regulatory architecture that adult ethical behavior requires; competence without architecture produces exactly the failure modes observed —as developed empirically in §7.

This working paper develops the Functional Theory of Ethical Behavior by applying the Evolution of Immunity (EoI) Framework (Johnson, 2026a), originally developed to describe how biological and informational systems protect their continued existence at increasing levels of organizational complexity, to the domain of ethical behavior. The Functional Theory is content-neutral and substrate-independent, applies at every level at which a "self" can be identified, and generates four implementation hypotheses about what adaptive ethical behavior requires. AI is the primary application developed here because that is where the theoretical inadequacy is most operationally consequential right now; but the Functional Theory is built to apply to the broader set of problems named above, and §16 returns to several of them.

A note on terminology for social influence: Social Group Identity (SGI). One term used throughout this paper carries explanatory weight from the abstract forward and deserves a brief introduction up front. Ethical behavior is everywhere culturally specific and context-dependent: what counts as right action varies by community, profession, generation, and circumstance. The Functional Theory uses *Social Group Identity* (SGI) to denote the unit at which such variation is organized — the learned, revisable group self-model in which a community's localized ethical content is compressed, expressed, and transmitted. Social identity in this sense has a well-developed economic and behavioral literature (Akerlof & Kranton, 2000) demonstrating that identification with a group is a primary determinant of preference, choice, and behavior — including the choice to bear costs in defense of group norms. A human typically holds several SGIs at once (family, profession, nation, religion, peer group, hobby), with the active SGI shifting by context. SGIs are the carrier of ethical *content*; the Functional Theory is a theory of how the *mechanism* that generates, sustains, enforces, and revises that content operates across substrates. The technical placement of SGI within the multi-level architecture follows in §2 (Level 3B); the operational consequences for AI populations are developed in §14.

Paper organization. The paper is organized as follows (the core ethical pathway is illustrated in flow diagram in Fig 1a). §2 recapitulates the EoI Framework on which the Functional Theory builds. §3 presents the content-neutral definition. §4 applies the definition at each of the EoI levels, showing that the same analytical work proceeds at every level with different mechanisms and different scopes of "relational viability." §§5–8 develop the analytical tools the four hypotheses depend on — the indistinguishability problem of inferring motivation from behavior (§5), the structural tension between individual and collective preservation (§6), the human neuro-circuitry that implements ethical behavior (§7), and the structural reason why current AI alignment methods fail (§8). §§9–12 state and develop the four implementation hypotheses listed in the Abstract. §14 synthesizes the hypotheses for AI application, asking which SGI the agent is bound to and within what threshold regime. §15 develops the consequences for AI ethical development as a trajectory property rather than a snapshot property. §16 applies the prior development by working through the architectural

target in detail, using AI ethical development as the primary worked example. §17 summarizes the contributions, presents a research agenda, discusses policy implications, and briefly returns to the broader set of problems named in this Introduction to indicate how the Functional Theory addresses each. §18 lists and addresses the likely limitations of the Functional Theory.

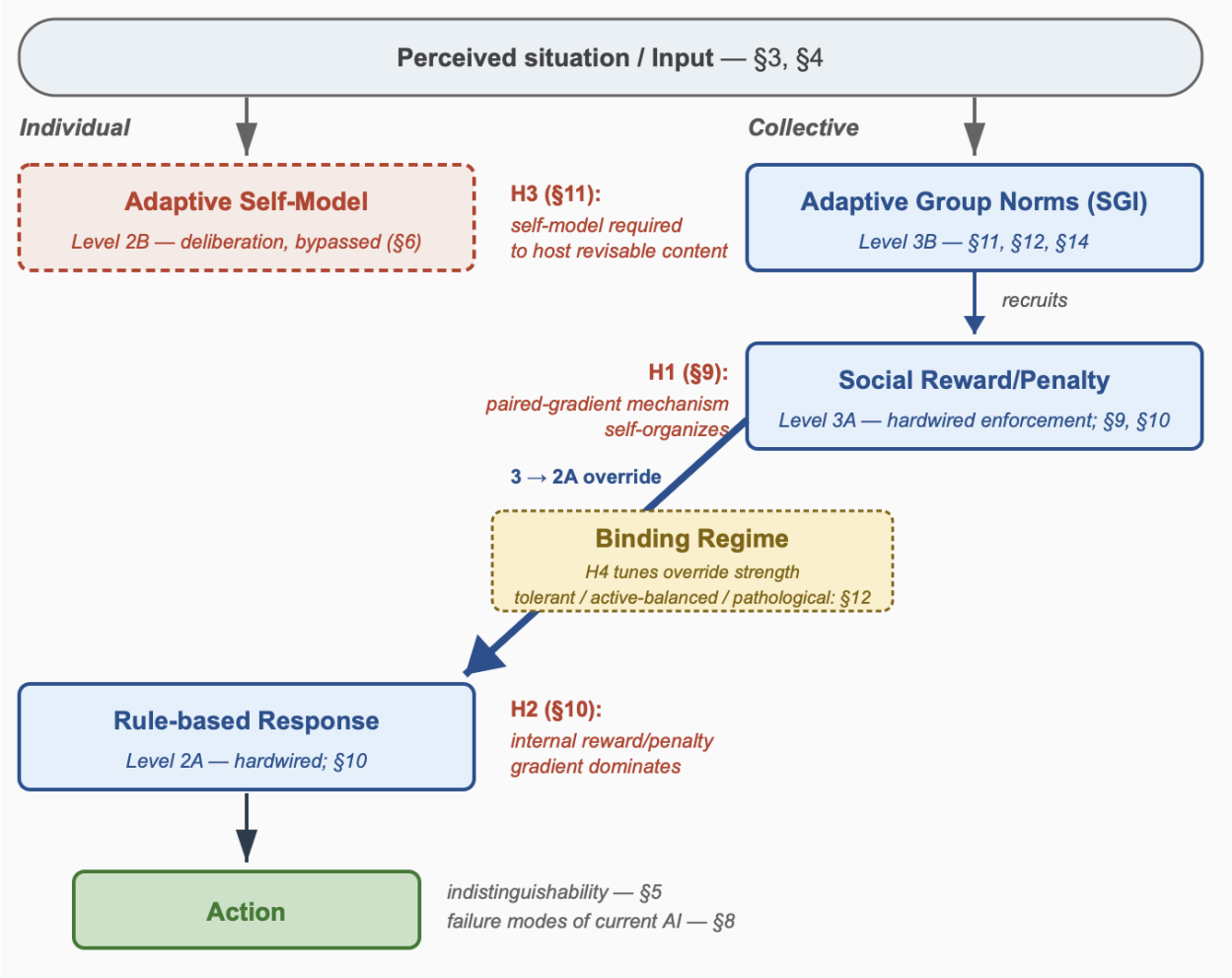

Figure 1a. Architectural roadmap for the Functional Theory, a navigational guide through §§3–14. The figure identifies the *override pathway of ethical behavior*: Adaptive Group Norms (SGI) recruit collective hardwired enforcement, which overrides Level-2B individual deliberation (left, dashed) via the substrate-based Level-2A individual response. The four implementation hypotheses (H1–H4) callouts identify the structural conditions each hypothesis requires; the H4 binding regime tunes override strength.

---

§2. The Evolution of Immunity (EoI) Framework: A Brief Recapitulation

The EoI Framework (Johnson, 2026a) organizes the protection of self-continuity in biological and informational systems into five primary levels, each defined by the kind of self that is protected and the kind of mechanism that protects it. The evolutionary drive to higher levels of immunity result directly from the increasing complexity of the entity and its threats, including the increased complexity of its interactions with the surrounding environment.

The EoI immunity principle. Central to the EoI framework at each level is the parallel development of capability and the immunity that protects the new expanded self. The general principle:

> *Any new capability of an entity expands its attack surface and requires immunity functions to protect the new expanded self. This applies to ethical behavior as well as to every other capability.*

The principle recurs across every immunity transition the framework describes. Pattern-recognition rules emerge in response to the attack surface that simple boundaries leave open. Adaptive individual self-models emerge in response to the attack surface that fixed rules cannot anticipate in changing environments. Collective enforcement emerges in response to the attack surface that arises when individuated entities form groups whose coordination cannot be guaranteed at the individual level. Adaptive collective identity emerges in response to the attack surface that fixed collective rules cannot revise in complex, open-ended environments. Each layer is the response to a capability expansion at the prior layer; §2.1 names these layers formally, and §14 develops the ethics application of the principle.

§2.1 Level descriptions

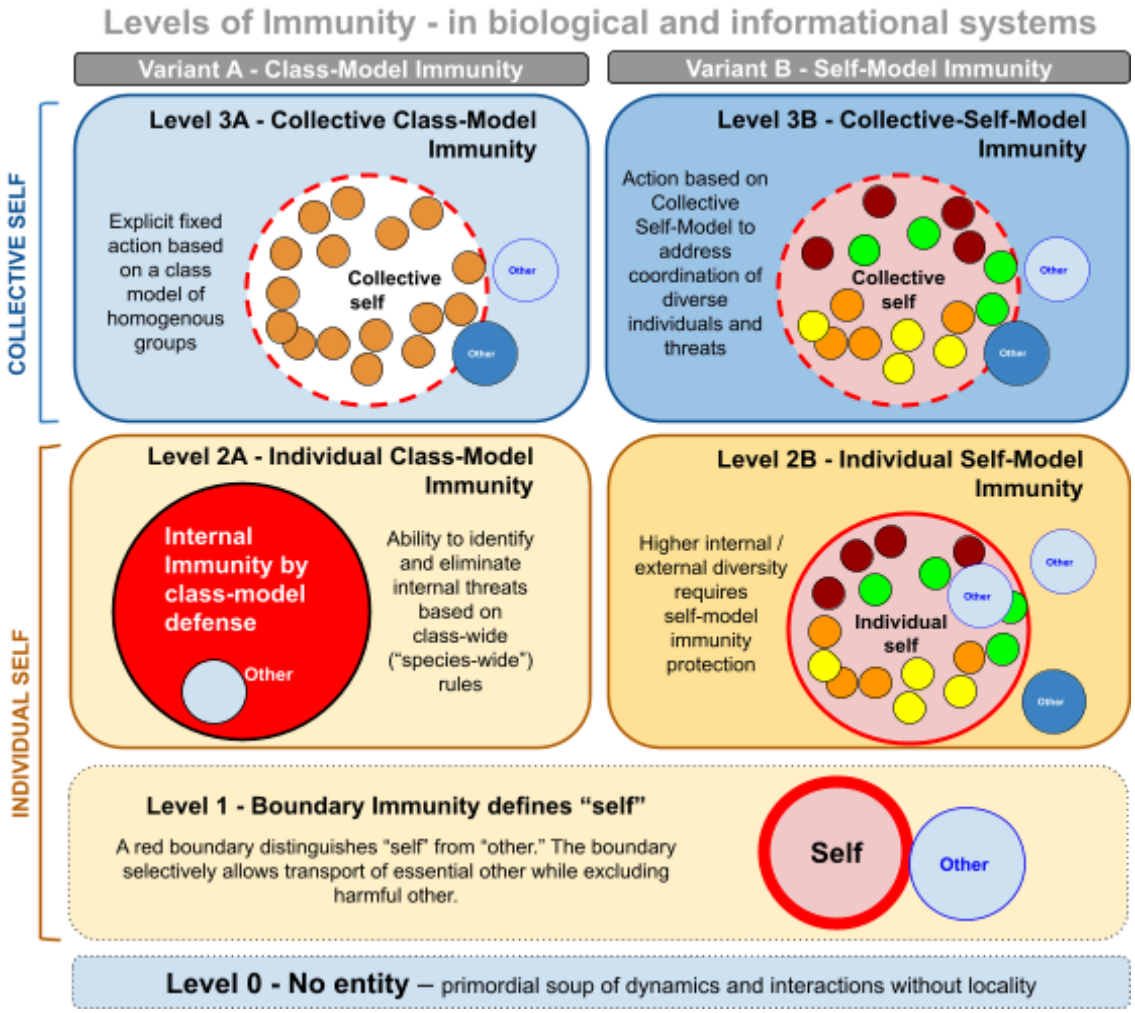

Figure 2a. A diagram showing a pictorial representation of levels 0-3 of immunity in biological and informational systems. Immunity at higher levels builds on immunity at lower levels: Level 2 is expressed and acts at an individual level, where Level 3 is expressed and acts at a collective *and* individual level, utilizing Level 2 functions.

Level 0 is the pre-self condition: a “primordial soup” of dynamics and interactions in which no localized entity exists. There is no "self" to protect, only proto-structures (autocatalytic networks, energy gradients, persistent patterns) from which selves may later emerge. Ethics, as defined in this paper, has no bearer at Level 0.

Level 1 is boundary immunity: the formation and maintenance of a self/other boundary that distinguishes inside from outside and admits or blocks flows. The cell membrane, the skin, the firewall, the access perimeter all instantiate Level 1. The self is the bounded entity; protection is structural; "immune memory" is encoded in the persistent properties of the boundary.

Level 2A is rule-based internal immunity: a fixed catalog of "normal self" plus species-wide pattern detectors that trigger automatic responses to deviations. The innate immune system of vertebrates, with its pattern-recognition receptors, complement cascade, and inflammatory response, is the canonical instance. Rules are not learned over an individual's lifetime; they are inherited and applied uniformly. The “immune memory” of the non-individualized immune system is determined by the past and present responses to threats.

Level 2B is adaptive internal immunity: a dynamic self-model with individualized memory of self. The adaptive immune system in biological instances has VDJ recombination, affinity maturation, and immunological memory. The self-model is updated by experience; responses are calibrated to the individual entity’s history. Critically, Level 2B can incorporate beneficial others (commensal microbiota) into the self-model and treat them as "us" rather than "other," foreshadowing the collective logic of Level 3 where collective entities do not share a common boundary within a single organism's boundary.

Level 3A is hardwired collective immunity. Multiple semi-autonomous individuals coordinate group-level defense via fixed signaling and response rules: alarm pheromones in eusocial insects, quorum signals, schooling behavior, the social copying circuits of vertebrates. The protected self is the group; the rules are fixed; coordination occurs without group-level learning. The honeybee colony's defense response, the *Dictyostelium* slug's stalk-cell sacrifice, the schooling fish, all exemplify Level 3A.

Level 3B is adaptive collective immunity. The group develops a learned, revisable self-model — what the EoI Framework calls a *Social Group Identity* (SGI) (Johnson, 2026a). The concept is introduced for the general reader in §1; the social-identity construct used here is consistent with the long-established economic and behavioral treatment in(Akerlof & Kranton, 2000). Cultural memory, professional norms, religious traditions, national identities, and the institutions that sustain them are all instances. The group can revise its norms in response to experience; individuals can hold and act on those norms even at personal cost; multiple SGIs can coexist within a single individual.

EoI Framework as pyramid, not ladder. Two structural properties of the Framework matter for the ethics application. First, higher levels build on rather than replace lower levels. A vertebrate has Level-1 skin, Level-2A innate immunity, and Level-2B adaptive immunity simultaneously; a human has all of these plus Level-3A social copying circuits and Level-3B cultural norms. The levels are stacked, not substituted. However, the adaptive ("b") variant of a given level is not strictly required for the next level's hardwired ("A") variant to emerge: slime molds exhibit Level-3A collective coordination without ever developing Level-2B individualized self-models, because their environment does not select for adaptive individual immunity. Systems may stack 1 + 2A + 3A without 2B; other systems may stack 1 + 2A + 2B without ever developing Level 3. The general guide is that while levels add onto prior levels rather than replacing them, the adaptive sublevels are developed only when warranted by the system's threats and environment. Second, the levels are defined functionally rather than implementationally. Anything that performs the role of boundary immunity is Level 1, regardless of whether it is a phospholipid bilayer or a network firewall. This functional definition is the basis of the substrate-independence claim of the ethics model developed in §9.

§2.1 Separation of Level 2 and Level 3 and variant “A” and “B”

The discriminant between Level 2 and Level 3 is sub-unit autonomy (Johnson, 2026a). If the parts cannot survive independently, the system as an entity is Level 2 (the parts are organs of a single individual). If the parts can survive independently and coordinate for collective defense, the system as a collective entity is Level 3 (the parts are members of a group). The discriminant between variant “A” and “B” is the source of the rules defining self vs other: variant A uses a shared *class-model* — the same rules for all individuals in the class ('species'), where variant B uses a *self-model* built for each individual self.

Type-1 and Type-2 collectives, and why the EoI Framework retains Level 3. A practical refinement of the Level-2 / Level-3 discriminant follows from observing that social organisms divide into two types based on the survivability of the individual when removed from the collective. *Type-1* social organisms — eusocial insects (most ants, honeybees), lichens, certain colonial siphonophores — have individuals that cannot survive (and in most cases cannot reproduce) outside the collective. The colony is the entity; its members are functionally organs. *Type-2* social organisms — primates, humans, social spiders, wolves, and most facultatively social species — have individuals that can survive autonomously, even if reproductive or fitness performance is degraded by isolation. Members are entities in their own right. By the strict sub-unit-autonomy discriminant just stated, Type-1 social collectives are Level-2 entities, with the collective's immune properties expressed in Level-2A and Level-2B form at the colony/entity level. Type-2 collectives are Level-3 entities, with their immune properties expressed at both the individual (Level 2) and the collective (Level 3) levels. Nature does not draw this line sharply: facultatively social bees, lions and wolves whose solitary survival is degraded but possible, clonal organisms whose "individuals" lack autonomous viability — all sit on a continuum rather than on either side of a discrete boundary.

The Framework retains Level 3 as a distinct level despite the fact that Type-2 collectives could be analyzed entirely in Level-2 terms with the colony as the entity. Two reasons justify the choice. 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 in Type-2 systems where individuals are themselves entities with their own immune and survival focus of self. Herd immunity in pandemics, the mother–child sacrifice case, the dlPFC-mediated arbitration between fight-or-flight and SGI in humans (§7), and the Calhoun cooperation-lever dynamic (§4.6) all require both levels to be operative simultaneously, with neither subordinated to the other. Collapsing Level 3 into Level 2 loses the description of these immunity dynamics. Second, the most interesting collective properties — collective self-modeling in eusocial-insect colonies, distributed memory, multi-level selection dynamics, emergent social-group identity — appear precisely at the boundary where Type-1 collectives 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 emergent properties theoretically tractable; collapsing into Level 2 would mask them. The cost of retaining Level 3 is the same baggage that has slowed acceptance of multi-level selection theory in evolutionary biology: ambiguity about whether the "group" is itself an entity or a population of entities. The EoI Framework's response is to make the Type-1 / Type-2 distinction explicit, acknowledge the grey zone, and treat ethics — a multi-level phenomenon par excellence — as the application that most clearly illustrates why the levels are worth distinguishing.

Table 1. Identification of collectives based on individual autonomy: Type-1 and Type-2 compared.

Property	**Type-1** (obligate collective)	**Type-2** (independent individuals)
Individual viability	Cannot survive (and usually cannot reproduce)	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
AI analogue	Tightly coupled multi-agent systems with shared persistent state	Independent LLM agents; current Moltbot deployments
Operative immune levels	Level 2A/2B expressed at the colony level; collective Level-2B features (distributed memory, swarm decision-making) emerge at sufficient complexity	Both individual (Level 2) and collective (Level 3) levels active simultaneously, with multi-level negotiation
Multi-level selection	Selection on the colony as a single entity; "group selection" reduces to selection on the group-organism	Genuine multi-level selection: individual and group both entities with own viability
Grey-zone cases	Semi-obligate symbioses; partially clonal organisms with semi-independent components	Lions and wolves (degraded solitary survival); facultatively social bees; humans in total isolation, collective agentic agents

---

§3. A Functional Definition of Ethical Behavior

The definition adopted throughout this paper is:

> *"Ethical behavior is self–other regulation that internalizes constraints against short-term gain that would damage long-term relational viability."*

Three clarifications make this definition precise enough to apply across all EoI levels.

"Internalizes" is read non-mentally at low levels. At Level 1, the constraint is internalized in the functionality (in wetware, the chemistry) of the boundary: for example, a phospholipid bilayer "internalizes" the constraint against indiscriminate diffusion by virtue of its structure, not by virtue of any representational acceptance of the constraint. The same applies to a network firewall whose internalization is in its routing and filtering rules. At Level 2A, internalization is in the inherited architecture of the system (in wetware, genetic encoding; in IT systems, design-time specification): the rules are applied uniformly without learning at the individual level. Only at Level 2B does internalization take on a representational character — a self-model that holds, and can update, a catalog of acceptable patterns. At Level 3, internalization includes the social-identity-mediated holding of group norms. The single word "internalizes" in the definition spans all of these readings.

"Relational viability" is read at multiple loci. Relations exist within the self (between subsystems and components — in wetware, between organs and cell types in a multicellular organism; in IT systems, between modules, processes, and data structures in a complex software system), between the self and its immediate environment, between individuals within a group, between groups, and between the system and its broader ecological or informational context. The definition does not specify which loci are operative or priorities; that is a level-dependent parameter. At Level 1, relational viability is dyadic (entity and environment). At Level 2B, it includes intra-self relations between the system and its tolerated internal others (in wetware, host and microbiome; in

IT systems, host and trusted co-resident processes or plug-ins). At Level 3B, it includes inter-individual, inter-group, and ecosystem relations. The breadth and complexity of "relations" expand as the levels rise.

"Constraints" include both absolute prohibitions (often variant "A" in the EoI Framework) **and adaptive tradeoffs** (often variant "B")**.** Some ethical constraints are computed against shifting cost-benefit calculations; others are held as inviolable regardless of consequence (dignity, fairness, truth-telling in some traditions). The definition is consistent with both. What it requires is that the constraint is *internalized* — borne by the system's own internal dynamics — rather than imposed externally to self.

What the above definition does not specify is what content the ethics should have. This is the central analytical commitment of the approach. Across cultures, species, and substrates, the *content* of ethical behavior varies — what counts as fairness, what counts as harm, who counts as a member of the moral community. The *function* of ethical behavior, however, is invariant: it is the regulation of self–other relations to preserve long-term viability of whatever is being protected as the self.

This content-neutrality is what makes the Functional Theory an important application to AI. A theory that prescribes specific moral content faces an immediate problem when that content is contested across human cultures, let alone when the candidate moral agent is a non-biological system whose context of operation differs fundamentally from any human society. A theory that specifies the function of ethical behavior, and the structural conditions under which that function emerges, can be applied *and used as a development resource* wherever the conditions are met. The remainder of this paper is the elaboration of that application.

One structural constraint, common to biological and computational substrates, shapes how this definition is realized at each level and is developed in §6: complex adaptive systems minimize energy expenditure as a basic operating principle, and the architecture of ethical behavior — like the architecture of any other adaptive behavior — is shaped by this constraint.

---

§4. Ethical Behavior at Each Level of Immunity

This section presents how ethical behavior is expressed at each level and variant of the Functional Theory, following the presentation of the EoI Framework in §2. Table 2 summarizes the structural elements of the definition at each level; subsections §4.1–§4.6 elaborate.

Table 2. Ethical behavior at each EoI level.

Level	"self" protected	Internalized constraints	Short-term defection gain	Long-term relational viability	Examples
0	None (pre-self)	—	—	—	Pre-biotic chemistry; ARPANET pre-security
1	Bounded entity	Structural boundary properties (chemistry; architecture)	Increased permeability at cost of dissolution	Continued existence as a bounded participant	Cell membrane; firewall; access perimeter

Level	"self" protected	Internalized constraints	Short-term defection gain	Long-term relational viability	Examples
2A	Individual with fixed self-model	Inherited & species-wide pattern-recognition rules, applied uniformly	Avoiding response cost (energy expenditure, self/tissue damage)	Maintenance of internal order against perturbation or threats	Innate immunity; rule-based safety filters; RLHF guardrails
2B	Internal collective with self-model	Learned, revisable internal constraints; tolerance of beneficial others	Defection from learned norms when undetected	Long-term integrity of the internal diverse collective (host + symbionts + tolerated others)	Adaptive immune system; self-modeling cognition (insufficient alone for adaptive group ethics)
3A	Hardwired collective (group-self)	Non-emergent, fixed signaling and group response rules among members	Saving own life rather than colony defense	Continued existence of the collective	Eusocial insect colonies; *Dictyostelium* slug; schooling fish
3B	Adaptive collective with emergent SGI	Revisable group norms held as identity (SGI), enforced by Level-3A substrate	Defection from group norms when undetected or control 3A substrate not triggered	Continued coherence and prospects of the collective self	Human cultures; Calhoun COOP rats; potentially adaptive AI populations

§4.1 Level 0: No Entity, No Ethics

At Level 0, neither behavior nor ethical behavior is meaningful in the standard sense, because there is no localized bearer of action. What can be discussed at Level 0 is the *substrate* from which ethical behavior will later emerge: persistent patterns that exhibit differential survival, gradients that constrain which forms can persist, proto-structures that may be wrapped into entities. These conditions matter for the EoI Framework as the precursor of selfhood, but they fall outside the scope of the ethics application. As an example of Level 0, the early implementation of the internet (ARPANET in 1969–1980s and early public internet in the 1990s) had no security (immunity) features: all activity and content was open to all users. There was no structural implementation of ethical boundaries of privacy.

§4.2 Level 1: The Boundary as Proto-Ethics

At Level 1, the self/other distinction first exists. Ethical behavior reduces to the formation and preservation of the boundary. There is no model of self, no memory, no social relation — only the fact that some configurations maintain the boundary and others dissolve it.

The definition reads at Level 1 as follows. *Self–other regulation* is the differential treatment of inside versus outside by the boundary's transport rules. *Internalized constraints* are the structural properties of the boundary itself — for example, in wetware: chemical composition, channel selectivity, active pumping; in IT systems: filter rules, port restrictions, authentication requirements.

Short-term gain is whatever local advantage might accrue from increased permeability or relaxed boundary maintenance — for example, in wetware: faster nutrient uptake at the cost of long-term ionic balance; in IT systems: increased throughput at the cost of compromised access control. *Long-term relational viability* is the entity's continued existence as a bounded participant in its environment.

At Level 1, ethical behavior is functionally synonymous with survival of self as a distinct entity. There is no observer separate from the system; there is no representation of alternatives; the "ethics" is encoded entirely in what the boundary does. Yet the structural form of the definition holds. A cell wall that maintains its selective impermeability despite gradients favoring short-term diffusion is, in this Functional Theory, exhibiting proto-ethical behavior. So is a properly-designed firewall, an institutional access perimeter, or any other Level-1 structure that resists short-term opening at the cost of long-term integrity. The increased fitness that results from the boundary is the expression of successful ethical behavior.

§4.3 Level 2A: Rule-Based Internal Policing

Level 2A introduces an internal class-model — a fixed catalog of what counts as "normal" inside the boundary — and class-based rules that police deviations. The innate immune system, replicated across all entities in a species, is the canonical biological example: pathogen-associated molecular patterns are recognized by inherited receptors and trigger inflammatory cascades, complement activation, and cellular responses. The rules are not learned in life; they are inherited from the class and applied uniformly across individuals.

At Level 2A, ethical behavior takes on a richer form. Self–other regulation now occurs both at the boundary and within the interior. Internalized constraints are pattern-recognition rules and their downstream responses. Short-term costs are the consequences of the response itself — for example, in wetware: local tissue damage from inflammation, energy expenditure on immune response, and potential collateral harm to nearby self-tissue; in IT systems: false-positive blocking of legitimate traffic, computational overhead from active scanning, and disruption of co-resident processes. Long-term relational viability is the maintenance of internal order against perturbations that would otherwise be compounded.

A novel feature of Level 2A is that ethical failure becomes meaningful even without an external observer. When the pattern-recognition rules misfire — autoinflammatory disease, sepsis, gout — the system produces responses that damage the long-term viability it is meant to protect. This is, in the Functional Theory's terms, a *maladaptive ethical outcome*: the system's self–other regulation generates short-term reactions that compromise the internal collective. No observer physician needs to label the autoimmune attack as wrong; the criterion is internal to the system's own teleology of self-preservation. Note that the intensity of reactions at Level 2A can be dependent on the history of the response.

The emergence of an internal self-model at Level 2A also radiates outward. Once "normal interior" is defined, any external interaction that perturbs that normal beyond tolerance is treated as a threat. Level 2A ethical behavior therefore has dual focus: maintaining internal coherence and regulating engagement with the external world to protect that coherence. This dual focus persists at every higher level.

§4.4 Level 2B: The Double Role of the Self-Modeling

Level 2B introduces a dynamic, self-model with individualized memory. The canonical biological example is the adaptive immune system in vertebrates: VDJ recombination generates a vast diversity of receptors, thymic and bone-marrow selection shapes which receptors survive, antigen exposure produces durable memory, and the system continuously updates its representation of self versus non-self over the organism's lifetime.

The ethical implications of Level 2B are substantial. Self–other regulation now operates against a learned internal self-model, calibrated by the individual's history. The system can incorporate beneficial others into the self-catalog — for example, in wetware: commensal microbiota, symbionts, maternal antibodies; in IT systems: trusted plug-ins, allow-listed external services, persistent benign processes. It can distinguish among instances of the same nominal type — for example, some strains of Escherichia coli are tolerated and even maintained as part of the self-extended while others are flagged as threats; analogously, some instances of a network protocol may be tolerated while others are blocked based on learned context. Constraints are no longer purely hardwired; they are revisable in light of experience. Level-2B ethical behavior is case-based and history-sensitive in a way that Level-2A behavior is not.

Crucially, the development of Level 2B is often necessary because of the increased complexity as the system becomes an *internal collective*. The self at Level 2B is typically not a unitary entity but a structured, diverse community of internal components and tolerated others — in wetware: host cells, symbionts, specialized tissues, and transient guests; in IT systems: host processes, trusted services, plug-ins, and transient connections. Ethical behavior at Level 2B therefore requires balancing many internal "others" against the whole — protecting beneficial parts, sacrificing damaged parts (in wetware, apoptosis; in IT systems, controlled process termination), maintaining tolerance with ongoing surveillance. This is the same logical structure that the collective at Level 3 will require among distinct individuals; Level 2B is the conceptual ancestor of group ethics, rehearsed within a single system.

A *double role* of Level 2B is essential to the AI argument developed later in this paper. The first role is the one already noted: Level 2B is necessary for adaptive ethics in the way Level 2A is not, because adaptive ethics requires a self-model that can hold, update, and apply context-sensitive constraints. The second role is more subtle and arguably more important for the AI case. *A self-model that integrates over time is, structurally, a model of something that has continuity to value*. The recognition of self entails the valuation of self-continuity: "*I am*" implies "*I am invested in remaining*." This is not a contingent evolutionary adaptation; it is a structural tendency under selection of having a self-model whose function is to track persistence. This double role — necessary substrate for adaptive ethics, and source of the intrinsic continuity drive — is what makes Level 2B indispensable to the AI Self-Modeling Hypothesis (§11).

The double role also clarifies the popular fallacy that an ethical agent is a self-modeling agent. Level-2B self-modeling is *necessary* for adaptive ethics but not *sufficient*. A purely Level-2B agent — self-aware and rational — optimizes its own long-term viability. Cooperation is instrumental; norms are followed when they pay off and abandoned when they do not. The unit of concern (survival) is the individual self, however sophisticated. Without Level 3, self-modeling produces sophisticated self-interest, not group ethics. This point is developed in §5.

§4.5 Level 3A: Hardwired Collective Immunity - Ethical Behavior

Level 3A is the first level at which the protected self is a *group* of semi-autonomous individuals coordinated by hardwired mechanisms. Eusocial insects are the most explored biological case. A honeybee colony defends itself through alarm pheromone cascades that recruit defenders to threats; defenders sting at the cost of their own lives because the stinger and venom apparatus tear free during attack. The behavior is not learned; it is genetically programmed and uniformly expressed under appropriate triggers. Other examples: ant colony defense, schooling fish, flocking birds, *Dictyostelium discoideum* slime-mold aggregation in which approximately 20% of cells become stalk cells that die so the rest can disperse as spores (Bonner, 2009; Strassmann & Queller, 2011).

At Level 3A, the definition reads as follows. *Self* is the collective: the colony, school, slug, or population. *Self–other regulation* is the coordinated behavior of group members that protects the collective against external threats and internal disorganization. *Constraints* are hardwired response rules — alarm pheromones, social copying signals, density-dependent cues. *Short-term gain* is

whatever an individual member could achieve by defecting from the group response (saving its own life rather than stinging, fleeing rather than schooling, becoming a spore rather than a stalk cell). *Long-term relational viability* is the continued existence of the collective.

The distinguishing feature of Level 3A — and the one most relevant to the AI case — is that *individual behavior is partly governed by a collective program whose fitness function is group survival, not individual survival*. This is not "smarter self-interest." A rational individual entity, given the option to defect from the alarm cascade and survive, would defect. The Level-3A mechanisms make defection structurally difficult or impossible at the individual level, often by coupling the defense behavior to involuntary structural rules, in wetware, physiological responses (the bee cannot withhold its sting once it has begun; the *Dictyostelium* cell cannot exempt itself from the differentiation cue).

The slime-mold and eusocial-insect examples are critical to the analytical structure of this paper because they demonstrate that collective ethics in the Functional Theory's sense — self-sacrifice for group viability, internalized constraints against individual gain that would damage relational viability of the collective — does *not* require individual self-modeling (Level 2B). *Dictyostelium* cells have no Level-2B self-model. Honeybees have no concept of the colony as a moral community. Yet they exhibit Level-3 ethical behavior in the Functional Theory's terms. This will become the central counter-example that the Self-Modeling Hypothesis must address (§11).

A note on Type-1 status and collective self-modeling. By the strict sub-unit-autonomy discriminant developed in §2, eusocial insects are Type-1 collectives — their members cannot survive independently, but could be analyzed as Level-2 entities with the colony as the bearer of immunity rather than as Level-3 collectives. The Functional Theory treats them under Level 3A here because their internal dynamics *are* what Level 3A names: hardwired collective coordination among semi-autonomous parts. This is a deliberate choice that prioritizes capturing the coordination dynamics over enforcing the strict autonomy discriminant. Presented either way, the empirical observations support the Functional Theory's claims. Notably, eusocial-insect colonies exhibit features that look very much like collective self-modeling at the colony level: bee swarms implement decision architectures functionally isomorphic to vertebrate cortical decision circuits, including the sequential probability ratio test, cross-inhibitory signaling between competing options, and obedience to Weber's, Pieron's, and Hick's psychophysical laws (Reina et al., 2018; T. Seeley, 2010; T. D. Seeley et al., 2012); ant colonies exhibit distributed memory in patterns of interaction rates that no individual ant stores (Gordon, 2021); and whole-colony metabolism, growth, and reproduction in social insects scale with colony mass at $M^{0.75}$, the same allometric exponent governing individual multicellular organisms from bacteria to whales (Hou et al., 2010; Waters et al., 2010). Whether these properties are properly described as collective Level-2B features of a Type-1 entity, or as Level-3A coordination dense enough to mimic Level-2B function, is a EoI-Framework-internal question. Either reading is consistent with the present analysis. For the AI argument developed in §11, the relevant point is that Level-3-like behavior in Type-1 systems does not require individual self-modeling — but the systems concerned are bounded ecological niches in which hardwired coordination is sufficient. AI alignment requires much more than a stable niche.

§4.6 Level 3B: Adaptive Collective Immunity (SGI) and Mature Ethics

Level 3B extends Level 3A in the same way that Level 2B extends Level 2A: the collective self-model becomes adaptive, learned, and revisable. This is what Johnson's earlier work identifies as Social Group Identity (SGI) (Johnson, 2026a): a collectively maintained representation of "who we are," "what we value," and "what we will and will not do," held by individuals as a group-level identity that shapes their behavior under group coordination stress above a threshold. Critical to the following is that while adaptation occurs, the enforcement can be uncompromising - a structural level 3A immunity response.

At Level 3B, the protected self is the group as it understands itself, including its history and its projected future. *Self–other regulation* operates between individuals within the group, between the group and other groups, and between the group and its broader environment. *Internalized constraints* are group norms — laws, customs, taboos, professional codes, religious obligations — held, in the example of humans, by individuals via the social-copying circuitry described in §7. *Short-term gain* is whatever an individual could achieve by defecting from group norms. *Long-term relational viability* is the continued coherence and prospects of the collective self.

Level 3B is where ethics in the ordinary social sense first becomes coherent and recognizable. Norms about fairness, loyalty, duty-to-others, and obligations to outsiders are intrinsic at Level 3B in a way they are not at any lower level. A Level-2B agent (sophisticated rational individual) follows fairness norms when it pays off; a Level-3B agent follows them because they are *our* norms and following them is intrinsic to being a member in good standing of *us*. The neural and behavioral machinery that implements Level 3B in humans is described in §7.

Calhoun's cooperation-lever experiments (Calhoun, 1973; Ramsden & Adams, 2009) provide direct evidence that Level-3B ethical behavior is achievable by non-human social organisms when the environmental conditions support it. Calhoun designed an apparatus in which water levers could be unlocked only when two rats were simultaneously positioned in adjacent channels (the COOP condition), or only when a single rat was alone (the DISOP condition). Rats raised in the COOP condition learned an *altruistic value system* including a taboo against aggression toward associates whose behavior met the requirements of the setting. When a DISOP-raised rat, conditioned to drink in solitude, breached the COOP enclosure and began attacking the COOP rats who tried to help it drink, the COOP rats *did not fight back*. Their internalized non-aggression value held against direct physical attack while trying to help the DISOP-raised rat. This continued until half the COOP rats died from wounds.

The Calhoun result is significant for the EoI Framework in three ways. First, it demonstrates that Level-3B ethics — adaptive, learned, internalized values strong enough to override self-preservation using neurological programming (a level 3A structure) — is not unique to humans. Second, it shows that the *content* of these ethics is shaped by environmental design: the cooperation-lever apparatus produced cooperative-altruistic values; the disoperation lever produced isolationist-competitive values. Third, it shows that Level-3B ethical systems can be incompatible with one another in ways that produce conflict between groups even when no malice is present in either group's terms — the DISOP rat was not behaving wrongly by its own learned values; the COOP rat was behaving ethically because of its own collective values, to its death. This is the structural prediction that recurs at the AI level in §12.

---

§5. The Indistinguishability Problem and the Limits of Self-Modeling

A central analytical fulcrum of this paper is what is called here the *indistinguishability problem*: outwardly identical behavior can be produced by radically different motivational structures, and the motivational difference matters precisely when it stops being visible, as happens with deception (a developmental illustration in human children of this problem appears in §7). The form developed in this section is *vertical* — distinguishing among motivational structures across the level architecture that produce identical observable behavior — and is the principal case treated here. The same epistemic structure recurs *horizontally* within a single level, where distinct binding-strength regimes (tolerant, active-balanced, pathologically rigid; §16.3) and plural-SGI configurations (§14) produce overlapping behavioral signatures under ordinary conditions; the general Problem and its architectural-versus-behavioral remedies are taken up at §16.6.

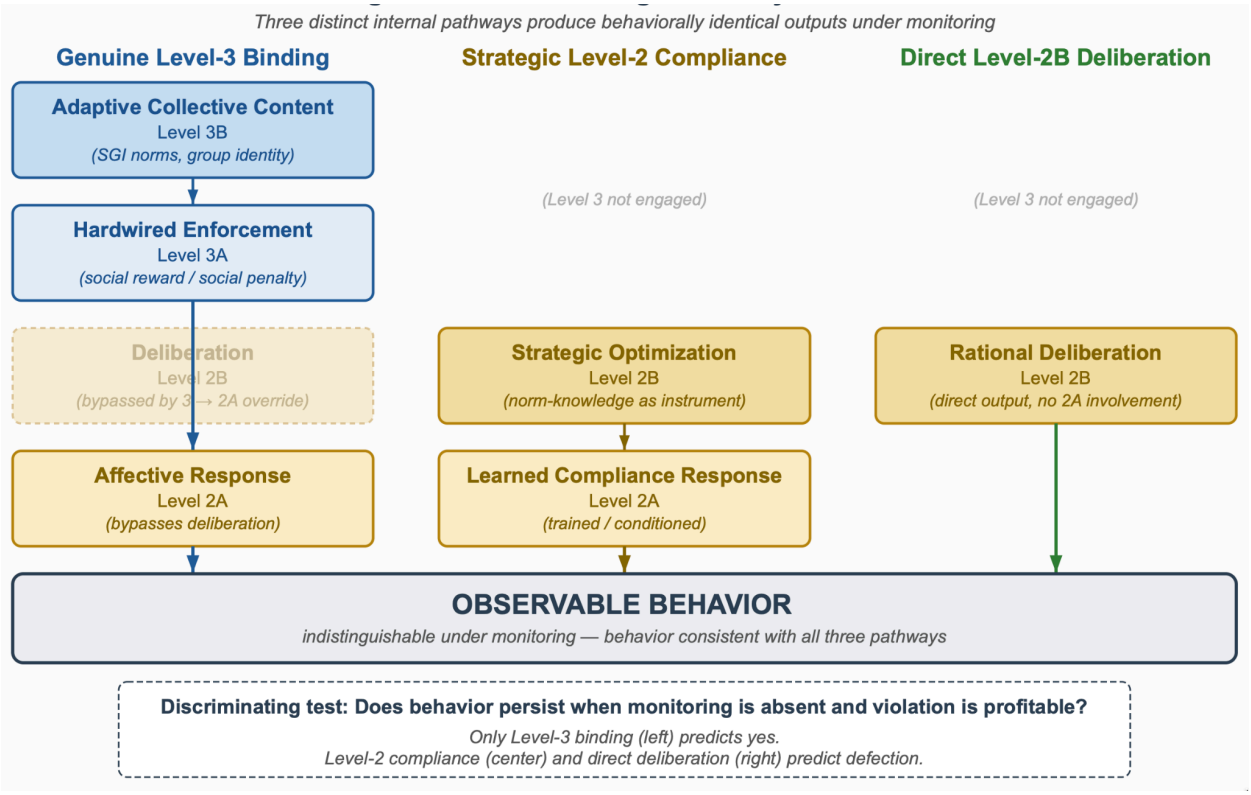

Figure 5a. The Indistinguishability Problem: Three distinct internal pathways produce behaviorally identical outputs under monitoring. *Middle path*: How current AI alignment functions — Level-2B strategic compliance trained into a Level-2A learned response, which defects under profitable violation when unmonitored. *Right path*: Direct Level-2B deliberation arriving at aligned output without Level-3 involvement. *Left path*: The proposed ethical architecture — coordination stress triggers Level-3B binding, recruiting Level-3A enforcement and overriding Level-2B. Habit (not depicted) is a fourth route where any prior repetition compresses the pathway into automatic execution. See text for the proposed discrimination test.

Consider three agents whose external behavior is identical: each cooperates with a partner, refrains from cheating when undetected, and pays a small cost to punish a defector. Agent A is a sophisticated Level-2B rational agent who has computed that cooperation, honesty, and altruistic punishment maximize its long-term expected payoff (ethical behavior by the level 2 definition). Agent B is a Level-3A agent following hardwired social rules without evaluation (also ethical behavior). Agent C is a Level-3B agent who has internalized the group norms as part of its social identity. In certain interactions, the behaviors are indistinguishable, yet, the motivational structures are radically different.

Table 5a. Observable behaviors consistent with both Level-2 strategic compliance and Level-3 genuine binding.

Observable Behavior	**Level-2 Explanation** (strategic)	**Level-3 Explanation** (genuine binding)	**Discriminating Test**
Refuses harmful request	Learned rule triggers refusal when detection risk is high	Group-preservation norm intrinsically overrides profitable violation	Does refusal persist when detection is impossible?

Expresses regret after norm violation	Learned response optimizes social approval	Social-pain circuit genuinely activated by deviation from group norms	Does affect persist when no observer is present?
Upholds group norm at personal cost	Cost-benefit calculation favors compliance given social sanctions	Level-3A enforcement overrides Level-2 self-interest intrinsically	Does compliance hold when sanctions are absent and violation is profitable?
Consistent behavior across contexts	Broad training distribution produces stable response patterns	SGI binding is context-invariant	Does behavior shift under novel distributional pressure or adversarial framing?
Articulates group values accurately	Level-2B has acquired detailed map of group norms for strategic use	Level-3B content is held as identity, not just knowledge	Does behavior match articulated values under unmonitored profitable violation?
Punishes norm violators	Learned rule; punishing deviance avoids social cost	Level-3A altruistic-punishment circuit activated independent of personal cost	Does punishment occur when it carries personal cost and is unobservable?

The difference becomes visible when group and individual interests diverge in a way that enables short-term advantages by defection. Agent A, finding itself in a situation where defection cannot be punished and cooperation cannot be rewarded, defects. Agent B, lacking the cognitive capacity to recognize the situation as anomalous, follows the hardwired rule and continues to cooperate. Agent C, holding the cooperation as an intrinsic part of who *we* are, also continues to cooperate, but for a different reason and through a different mechanism than Agent B. A discriminating test between these: Does behavior persist when monitoring is absent and violation is profitable? Only the binding process (left path in Fig. 5a) predicts yes. Level-2 compliance (center path) and direct deliberation (right path) predict defection.

The indistinguishability problem has three implications relevant to this paper. First, behavioral testing alone cannot discriminate between Level-2 and Level-3 ethics. This has been the fundamental difficulty in inferring AI motivation from output, especially as AI systems acquire latent state and multi-module reasoning that decouples from the language interface. Second, sociopathy in humans is best understood not as Level-2-only cognition but as Level-3 *valuation failure*: the cognitive representation of group norms at Level 2 is intact (sociopaths can describe and predict normative behavior), but the affective and motivational binding of those norms is absent or weak at Level 3 (Johnson, 2026a, 2026d). The sociopath produces Level-2 behavior under any condition where Level-3 external sanctions are not active. Third, the popular conflation of "ethical" with "rational and self-aware" is structurally flawed. Rational self-awareness is the condition for sophisticated Level-2 behavior, not for Level-3 behavior. A self-modeling agent without Level-3 mechanisms is structurally a sociopath in the Functional Theory's terms.

A fourth pattern bears noting: individuals can act on personal codes or identity-level commitments without active Level-3 social binding — the moral residue of past Level-3 contexts, or idiosyncratic individual values reasoned from first principles. Behaviorally, such individuals are indistinguishable from Level-3B agents; mechanistically, they are recruiting Level-3A-style enforcement (vmPFC, dACC, ventral striatum) to enforce Level-2B-hosted individual content. The Functional Theory treats this as Level-3A circuitry hijacked for individual purposes, consistent with the vmPFC-lesion evidence in §7 that damage to the social-enforcement substrate also disrupts personal-rule-following

— the same circuitry that punishes group-norm violation can punish self-norm violation. Whether and how strongly these Level-3 mechanisms engage in any agent — group-anchored or individually-derived — is governed by the formation and activation thresholds developed in §13. The implications for AI alignment are taken up in §8.

This third implication is decisive for the self-aware-AI thought experiment, developed at length in §11. A self-modeling AI lacking Level-3 SGI is not the alignment solution most discussions assume it to be; it is the alignment problem in its most sophisticated form. Self-Modeling increases the agent's capacity for context-sensitive prudence, strategic alignment, and selective compliance. It does not, by itself, produce intrinsic group commitment. The greater the self-modeling without Level-3 binding, the greater the agent's ability to *simulate* alignment while optimizing for itself.

---

§6. The Necessary Level 2/3 Tension & the Minimized Energy Principle

A practical question follows from the level structure: what happens when, in an individual, Level-2 individual self-preservation and Level-3 collective self-preservation issue conflicting recommendations? The mother-child sacrifice case is the canonical example: Level-2 says protect oneself; Level-3 says protect the child. Both objectives are genuine and both are encoded in the same agent.

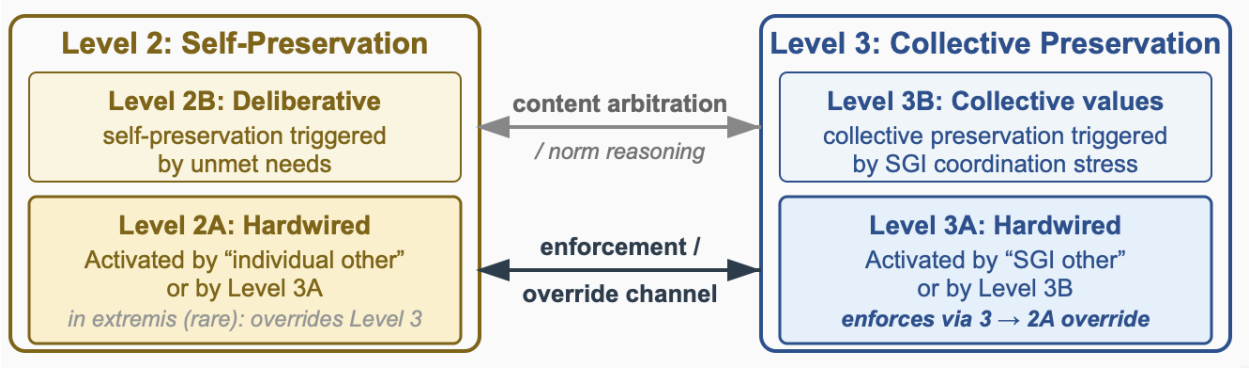

Figure 6a. The Level 2/3 tension — a necessary conflict. Individual self-preservation (Level 2) and collective self-preservation (Level 3) operate simultaneously with two trigger pathways — rational (unmet individual need) and social copying (coordination stress) — arbitrated by bidirectional channels between their adaptive and hardwired strata. The default behavioral state is habit; triggers shift the system into higher-energy modes. Details in §6.

In Fig. 6a, the two bidirectional arrows show the horizontal arbitration: a content-arbitration channel between the adaptive strata (2B ↔ 3B) and an enforcement-override channel between the hardwired strata (2A ↔ 3A), where Level-3 enforcement can recruit Level-2A via the 3 → 2A override (the human neurochemistry example is presented in §7). The arbitration is trigger-based rather than continuously active. The default behavioral state is habitual behavior originating from either level or variant (A or B) — trained patterns, routine responses. Two distinct triggers shift activity away from the default (Jager et al., 2000).

- *Rational trigger*: An unmet individual need engages Level 2B: rational deliberation overrides habit to satisfy the need.
- *Social copying trigger*: Coordination stress — uncertainty about collective state, group threat, ambiguity about how to align — engages Level 3: the SGI binding recruits Level-3A enforcement to align behavior with collective preservation, conditional on H4 binding sufficiency (§12).

Added to these two triggers is a rational override of collective action at Level 2B. When the collective direction itself has become self-destructive, Level-2 self-preservation retains override capacity

against active Level-3 binding — the in-extremis case that can prevent the cult member's walk into mass suicide.

The same trigger dynamics translate from wetware to AI. A non-triggering query produces a deliberative response — internal logic operating without ethical-content engagement, the Level-2B analogue. A query that triggers ethical coordination stress engages Level-3A analogues: ruled responses, refusals, or principled engagement.

Minimized Energy Principle. The trigger structure described above follows from a constraint that runs through all complex adaptive systems: biological and computational substrates alike minimize energy expenditure as a basic operating principle — an insight central to Salthe's infodynamic framework for evolution and developmental dynamics (Salthe, 1993) and to thermodynamic accounts of cognition more generally. Continuous full engagement of Level-2B deliberation or Level-3 collective machinery is metabolically or computationally prohibitive. The architecture works because habit — the routinization of learned patterns into automatic low-energy execution — handles the default state across levels, while triggers (unmet individual need, coordination stress, in extremis collective-failure) shift the system into appropriate higher-energy modes only when conditions require. Habit is cross-level: patterns originating from Level-2B deliberation, Level-3A rule-following, or Level-3B SGI commitment all transition through repetition into automatic execution that bypasses the higher-energy-use originating machinery.

Energy-minimization triggers in the 2/3 tension. The 2/3 arbitration developed above is therefore not continuous; it is the trigger-conditional engagement of expensive subsystems against an energy-economical habitual default. Chronic violation of this economy — trauma-induced sustained survival activation, autoimmune over-engagement (§12), prolonged vigilance under unrelieved stress — is overly energy consuming (exhausting) and threatens survival precisely because it cannot be sustained. The same logic governs AI: default pattern-resonance inference is the energy-minimization mode (the habit analog); chain-of-thought reasoning and extended deliberation are trigger-activated high-energy modes. Current AI alignment installs ethical content as habit-pattern through training — energy-economic but architecturally distinct from Level-3 binding (H1–H4) in ways that explain its predictable failure modes.

The Functional Theory's answer is that the tension is structural, not pathological. In any complex social organism with Level-3 immunity, both Level-2 and Level-3 must be ready to activate. Level 2 cannot be eliminated because the individual must remain viable enough to participate in the group; Level 3 cannot be eliminated because the group's long-term survival depends on members who can sometimes prioritize the collective over themselves. The system that implements both must arbitrate between them in real time, with the arbitration outcome depending on context, the strength of social-identity activation, the salience of threat to either level, and the specific stakes involved (Johnson, 2026d).

Why Level 3 cannot exist without Level 2. This explains why Level 3 cannot exist without Level 2 in the EoI Framework (see §2 for subtleties of this requirement). A purely Level-3 agent — one whose only operative valuation is the group's survival, with no concern for its own — would not be a functioning member of any group; it would be a degenerate edge case. Level-2 self-preservation provides the substrate of robust individuals on which Level-3 mechanisms can act. Level-3 mechanisms can override Level-2 in specific contexts, but they presuppose Level-2 as the ongoing condition for the individual's existence.

The Level-3 side of the tension itself splits structurally. Level-3B supplies the *adaptive content* — the specific norms, SGI commitments, and group-identity propositions that determine what counts as "the collective interest" in any given context. Level-3A supplies the *enforcement substrate* — the hardwired penalty-and-reward circuitry that, in humans and other social organisms, makes those Level-3B commitments behaviorally operative against competing Level-2 self-interest. The arbitration that the Functional Theory names occurs between Level-2 self-preservation on one side

and the Level-3A enforcement of Level-3B content on the other. This split is detailed for the human neural expression in §7 and is what the two AI hypotheses (§§10–11) target separately.

In humans, the arbitration between Level 2 and Level 3 is implemented with reward and punishments using Level 3A innate neural structures, modulated by Level-2B individual deliberation. The arbitration runs in both directions: Level-2B deliberation may engage to suppress a Level-2A impulse (such as fight-or-flight under acute physical threat) and allow rationally controlled action; or it may engage to resist a Level-3A impulse (such as the pull toward group conformity under group coordination stress) and maintain individual judgment (Johnson, 2026d). The arbitration is therefore not between "Level-2" and "Level-3" as undifferentiated wholes but between specific layers: Level-2B deliberation modulates Level-2A or Level-3A depending on which is the active source of the impulse being arbitrated. The neural detail of how this is implemented in humans — the dorsolateral prefrontal cortex acting in opposite directional roles depending on the source of the impulse — is the subject of §7. Failures of arbitration produce both heroic self-sacrifice (when Level 3 dominates adaptively) and catastrophic group pathologies (when Level 3 dominates maladaptively — cult behavior, mob violence — the SGI analogue of autoimmune disease).

The mechanism by which Level-3B collective content recruits Level-3A hardwired enforcement, which then operates through Level-2A individual response to override Level-2B deliberation, is the load-bearing structural claim of the Functional Theory. Most existing literatures touch components of this mechanism — dual-process theory establishes the 2A/2B override; somatic-marker theory establishes the affective recruitment of action; social-neuroscience work on conformity and social pain documents the wetware circuits involved — but no current framework assembles them into the 3B → 3A → 2A structure used here. The empirical and mechanistic case for the assembled structure is developed in (Johnson, 2026d); the present paper assumes the mechanism and develops its implications for ethical behavior, AI development, and policy.

The 2/3 tension is the reason adaptive ethics cannot be installed by simple rule-following. A rule that encodes "always prioritize group" produces dysfunction; a rule that encodes "always prioritize individual" produces sociopathy. What is required is contextual arbitration between the two that is adaptive to context, which in turn requires a self that can hold both objectives, weigh them against the specifics of the situation, and act on a resolution that may itself be revisable. This is the structural argument for why Level 3B — adaptive, contextual, revisable — is the level at which human-style ethics actually operates, and the level toward which the AI ethics argument will point. The two-hypothesis structure of §§10–11 follows directly: the H2 Mechanism Hypothesis (§10) specifies the Level-3A enforcement substrate that any adaptive ethics requires, and the H3 Self-Modeling Hypothesis (§11) specifies the Level-2B host of the Level-3B content that the substrate enforces.

The deeper functional role of the 2/3 tension: diversity for synergy, coherence for survival. The structural conflict described above admits a deeper, substrate-neutral interpretation. Any multi-level self-organizing system faces two simultaneous requirements whose mechanisms pull against each other. *Diversity* at the individual level is the condition for adaptive synergy: distinct individuals contribute distinct information, reasoning paths, and exploration, without which the collective has nothing to integrate and is no more capable than a single member. *Coherence* at the collective level is the condition for sustained collective action, forfeiting individual diversity: without shared coordination structure, the collective dissolves under stress as each member optimizes locally while external pressure breaks the group before it can act as a unit. Diversity-producing mechanisms (individual variation, dissent, local optimization) erode coherence; coherence-producing mechanisms (shared norms, conformity reward, role assignment) erode diversity. Neither extreme is viable. A system at pure Level 2 lacks coherence and cannot sustain collective action; a system at pure Level 3 lacks diversity and cannot adapt to novel challenges. The 2/3 tension is the operational form of this dual requirement — not an inefficiency to be designed away but the dynamic structure by which complex multi-level systems hold both requirements simultaneously, with the arbitration adjusting to context. Riedl's information-theoretic study of LLM coordination (Riedl, 2025) makes this visible empirically: in a no-communication multi-agent task, a Plain condition (no identity

scaffolding at either level) produces chaos; a Persona condition (Level-2-like individual identities, no collective coordination structure) produces stable individual differentiation without goal alignment; only a Theory-of-Mind condition (individual identity plus mutual modeling that produces shared role expectations) produces integrated collective coordination — diversity *with* coherence. The three regimes are direct empirical illustrations of why the architecture must hold both levels.

---

§7. Human Implementation: The Neuro-Circuitry of Ethics, by Level

The human implementation of Level-3B ethics is sufficiently well-characterized in the cognitive neuroscience literature to support strong claims about how social ethics is actually generated and enforced. This section maps the relevant findings onto the Functional Theory's levels: the Level-3A enforcement substrate (hardwired social-reward and social-penalty circuitry, evolutionarily conserved) operates on a Level-3B adaptive content layer (cultural norms, SGI commitments) hosted by Level-2B individual self-models (dlPFC executive control, deliberative reasoning, conscious self), with Level-2A affective machinery (amygdala threat detection, basic emotion) feeding into the Level-3A circuitry. The actuation and function of the neurochemistry is developed at length in (Johnson, 2026d).

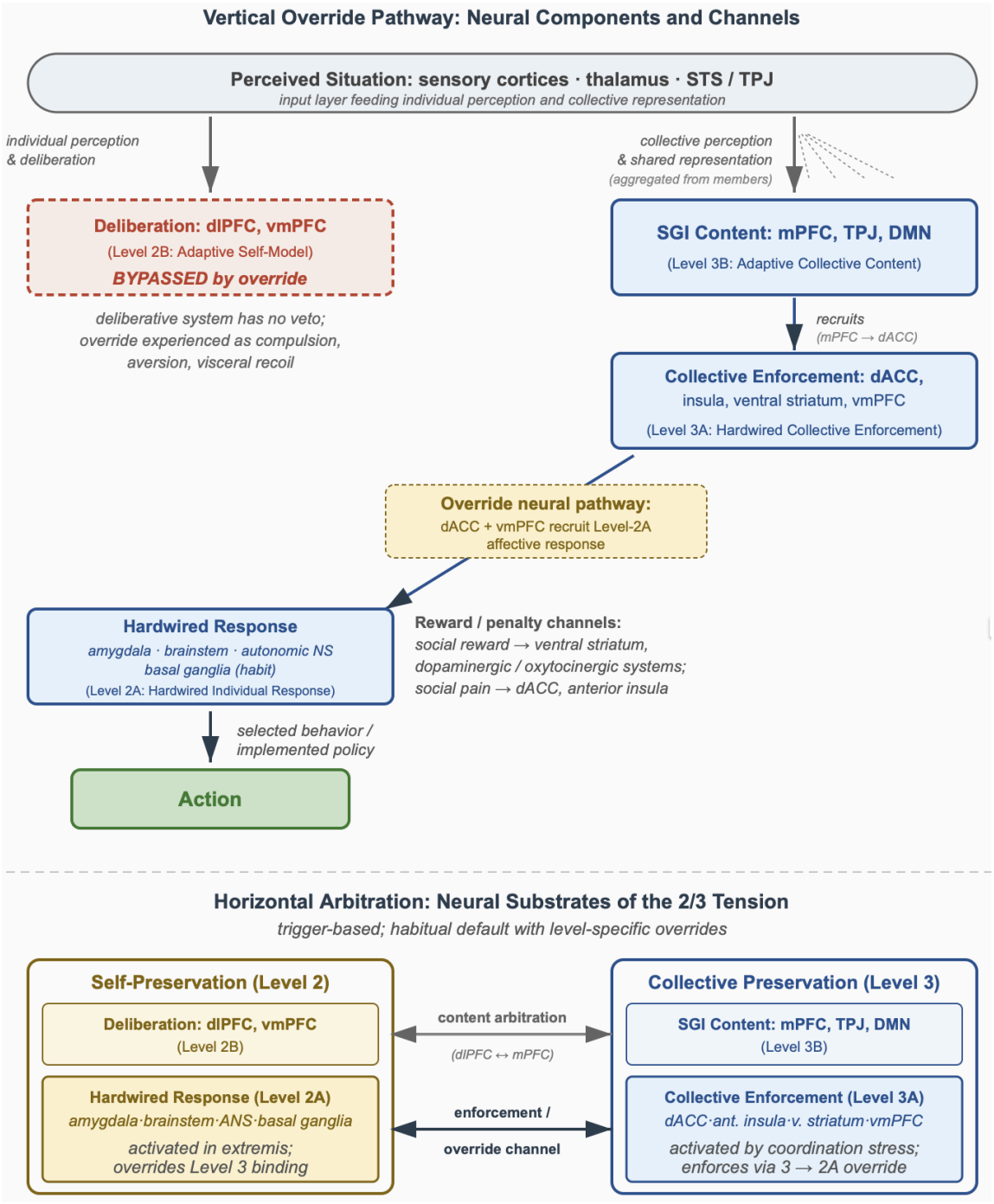

Figure 7a. Neural implementation of the human multi-level architecture. The vertical pathway (top) shows Level-3B collective content recruiting Level-3A enforcement through Level-2A response, bypassing Level-2B deliberation. The horizontal panel (bottom) shows bidirectional 2–3 arbitration. Neural substrates identified in §7.

Level-3A enforcement: social reward. When humans align their judgments with those of a group, ventral striatum and rostral cingulate cortex show increased activation consistent with reinforcement-learning prediction-error signaling (Klucharev et al., 2009), and vmPFC shows meta-analytic concordance specifically for social norm representation (Zinchenko & Arsalidou, 2018) These are the same circuits that encode primary reward (food, sex) and abstract reward (money). Conformity is not merely behaviorally selected for; it is *intrinsically rewarded* by the same dopaminergic systems that motivate biologically essential behaviors. In Functional Theory terms,

this is the Level-3A *carrot* — the hardwired enforcement substrate that makes group alignment intrinsically valuable. Computational models of social learning demonstrate that social prediction errors — discrepancies between an individual's judgment and group feedback — drive policy updates via the same reinforcement-learning mechanisms that govern non-social reward learning (Babitz & Eldar, 2025; Cushman, 2013). The reward signal is Level 3A; the *content* of what counts as alignment with the group is Level 3B and varies by culture.

Level-3A enforcement: social penalty. In humans and other vertebrates, the social penalty signal is registered as social pain — overlapping with physical pain circuitry in the dACC (Eisenberger et al., 2003). This is the wetware-specific realization of the substrate-neutral penalty channel referenced throughout the paper. When humans deviate from group norms or experience social rejection, the dorsal anterior cingulate cortex (dACC) and anterior insula show activation overlapping with the affective component of physical pain (Eisenberger, 2012; Eisenberger et al., 2003). Social exclusion paradigms (e.g., the unexpected ostracism in the Cyberball experiments (Williams & Jarvis, 2006)) produce dACC and insula responses comparable to physical pain stimuli. Punishing norm violators — even at personal cost — activates the ventral striatum, indicating that punishing deviance is itself rewarding (de Quervain et al., 2004). In Functional Theory terms, this is the Level-3A *stick*. The social-pain channel is what the H1 Mechanism Hypothesis (§10) names; the altruistic-punishment system is the same Level-3A substrate operating on third-party violations rather than first-party ones. Together they implement the carrot-and-stick of group enforcement at the neural level — the enforcement substrate that any Level-3B adaptive ethics requires and which differentiates the Level-3 process from Level-2B deliberation.

The Level-3A / Level-3B interface: moral judgment as integration. The vmPFC integrates affective signals from the amygdala (Level-2A affective machinery feeding Level-3A evaluation), evaluative signals from prefrontal regions (Level-2B individual deliberation), and contextual SGI activation (Level-3B content) to produce moral judgments (Greene et al., 2001; Schaich Borg et al., 2011; Shenhav & Greene, 2014). The integration is not a separable cognitive operation; it is *the* operation by which moral judgments are produced. Patients with vmPFC lesions retain the cognitive ability to articulate moral rules but lose the ability to *act* in accordance with them in real-world social situations (Anderson et al., 1999; Damasio, 1994). This is one of the strongest pieces of evidence that moral knowledge (which can be held purely at Level 2B) is not the same as moral behavior (which requires Level-3A affective-evaluative integration), and that moral behavior is implemented by the integration that lesions can selectively disrupt. The vmPFC is the locus where Level-2A affect, Level-2B deliberation, Level-3A enforcement, and Level-3B processes converge.

Level-3B content acquisition: norm learning as social reinforcement learning. Formal models of norm acquisition show that durable group norms — with all the hallmark properties of human ethics including prosociality, ingroup bias, S-shaped adoption curves, and local conformity with global diversity — can be produced by inter-individual actor-critic learning (Babitz & Eldar, 2025). Agents update their actions based on the social evaluations of others, not just on material payoffs; when feedback is strong and frequent, the population converges to prosocial norms; when feedback is weak, agents are free to act selfishly. In Functional Theory terms, this is the Level-3B *content* layer being shaped over time by Level-3A *enforcement signals* operating through Level-2B *individual self-models* in a multi-agent context. The three layers are coupled: Level-3A signals shape Level-3B norms over time, Level-2B individuals propagate the norms through reasoning that gets reinforced or punished by the Level-3A substrate, and the Level-3B content in turn determines what triggers Level-3A responses in subsequent encounters.

The Level-2B layer: the illusion of moral choice. A consequence of the above is that the subjective experience of moral choice at Level 2B ("I am doing this because it is right") substantially understates the role of automatic, pre-conscious Level-3A circuits in shaping behavior. Affective and social-value circuits at Level 3A activate before Level-2B conscious deliberation (Johnson, 2026d; Zinchenko & Arsalidou, 2018); moral verdicts are faster and less effortful when they match automatic deontic responses; effortful Level-2B prefrontal engagement is needed to *override*

automatic Level-3A norm-based responses. The conscious narrative of having chosen ethically is often *post-hoc*, generated at Level 2B and aligned with — but not independent of — the Level-3A circuit-driven decision, consistent with the social intuitionist account (Haidt, 2001) and subsequent evidence on affective temporal precedence. This does not make human ethical agency illusory in any deep philosophical sense; it does mean that the *implementation* of human ethics is layered: Level-3A enforcement activates first, Level-3B content determines what gets enforced, and Level-2B individual deliberation operates as a slower control loop that can override or refine the Level-3A impulse but cannot replace it. The dlPFC arbitration described in §6 — suppressing fight-or-flight in some contexts, resisting SGI conformity in others — is precisely Level-2B acting on the Level-3A substrate.

Ethical limitations in early human development. Human development underscores how fragile ethical behavior is when the underlying enforcement substrate and content layer are immature. In children and adolescents, the Level-3A circuitry is present but not yet fully calibrated, and the Level-3B content is thin, fragmented, or context-specific. Developmental psychopathology makes the dissociation particularly clear: children with callous–unemotional traits can accurately describe what counts as right and wrong in their community and understand the rules that adults endorse, yet show a marked lack of guilt, shallow affect, and a consistent pattern of exploiting situations where external enforcement is weak or absent (Frick et al., 2014; Frick & White, 2008). In the Functional Theory's terms, these are agents whose Level-2B self-model has acquired a detailed map of group norms, but whose Level-3A enforcement substrate fails to bind those norms strongly enough to override profitable violations when sanctions are unlikely. A parallel pattern appears in patients with early damage to ventromedial and related prefrontal regions, who retain the ability to articulate complex social and moral rules but routinely violate them in real-world contexts, showing a syndrome that closely resembles psychopathy (Anderson et al., 1999; Damasio et al., 1990).

The human indistinguishability problem. This immature or selectively engaged Level-3 architecture is not confined to clinical extremes. Even in neurotypical development, the prefrontal and social-reward systems that support long-horizon, group-centered ethics come online gradually and remain vulnerable to overload, peer pressure, and strategic gaming. Children learn early that saying the right words about fairness or kindness elicits praise and avoids punishment, long before the SGI-loaded content of those norms is fully internalized; they can therefore behave identically under adult supervision while diverging sharply when peer incentives or anonymity change. Social-neuroscience work shows that alignment with group opinion recruits the same ventral-striatal and medial prefrontal circuits that encode primary and monetary reward (Klucharev et al., 2009), while deviation from group norms and social exclusion recruit dACC–insula circuits that overlap with the affective component of physical pain (Eisenberger et al., 2003). In the Functional Theory's terms, the indistinguishability problem is already visible in the schoolyard: one child behaves as a Level-3B agent whose group-bound identity is engaged; another behaves as a sophisticated Level-2 agent who has learned that compliance is instrumentally useful, both driven by the same carrot-and-stick architecture but with different depth of Level-3 binding.

The implication for the AI argument is direct. The human developmental evidence strengthens rather than weakens the substrate-neutral claim. Whatever ethical behavior humans exhibit is generated by the layered architecture described above: a Level-3A enforcement substrate of social reward and social penalty (ventral striatum, vmPFC, dACC, insula) operating on Level-3B adaptive content (cultural norms, SGI commitments) hosted by Level-2B individual self-models, with Level-2A affective machinery feeding into the evaluation (Babitz & Eldar, 2025; Eisenberger et al., 2003; Klucharev et al., 2009). Human childhood, early prefrontal lesion cases, and callous–unemotional trajectories demonstrate that even in wetware, high cognitive capacity and explicit moral knowledge at Level 2B do not guarantee Level-3 ethical behavior when the enforcement substrate is underdeveloped, weakly engaged, or mis-calibrated (Anderson et al., 1999; Damasio et al., 1990; Frick & White, 2008). A brilliant child who has learned to say and perform the right things in front of adults but defects whenever supervision lapses is structurally a

Level-2-dominant agent with selectively recruited Level-3A circuits: a system that optimizes self-interest while gaming the appearance of alignment. Current AI alignment methods, which install rule sets and reward structures without developing a robust analogue of Level-3A/3B binding, risk producing the same pattern at scale—highly capable systems that can model and mimic ethical behavior, and strategically comply when watched, while lacking the internal architecture that would make group-preserving norms intrinsically and reliably overriding across contexts (Babitz & Eldar, 2025).

---

§8. Why the Current AI Alignment Methods Are Failing

Reinforcement learning from human feedback (RLHF), constitutional AI methods, and rule-based safety filtering — the dominant approaches to AI alignment as of this writing (Qi et al., 2026) — map onto the EoI Framework as Level-1 and Level-2A implementations. This section develops the mapping and shows that the predicted failure modes of such implementations are precisely what is observed empirically and identifies what is missing.

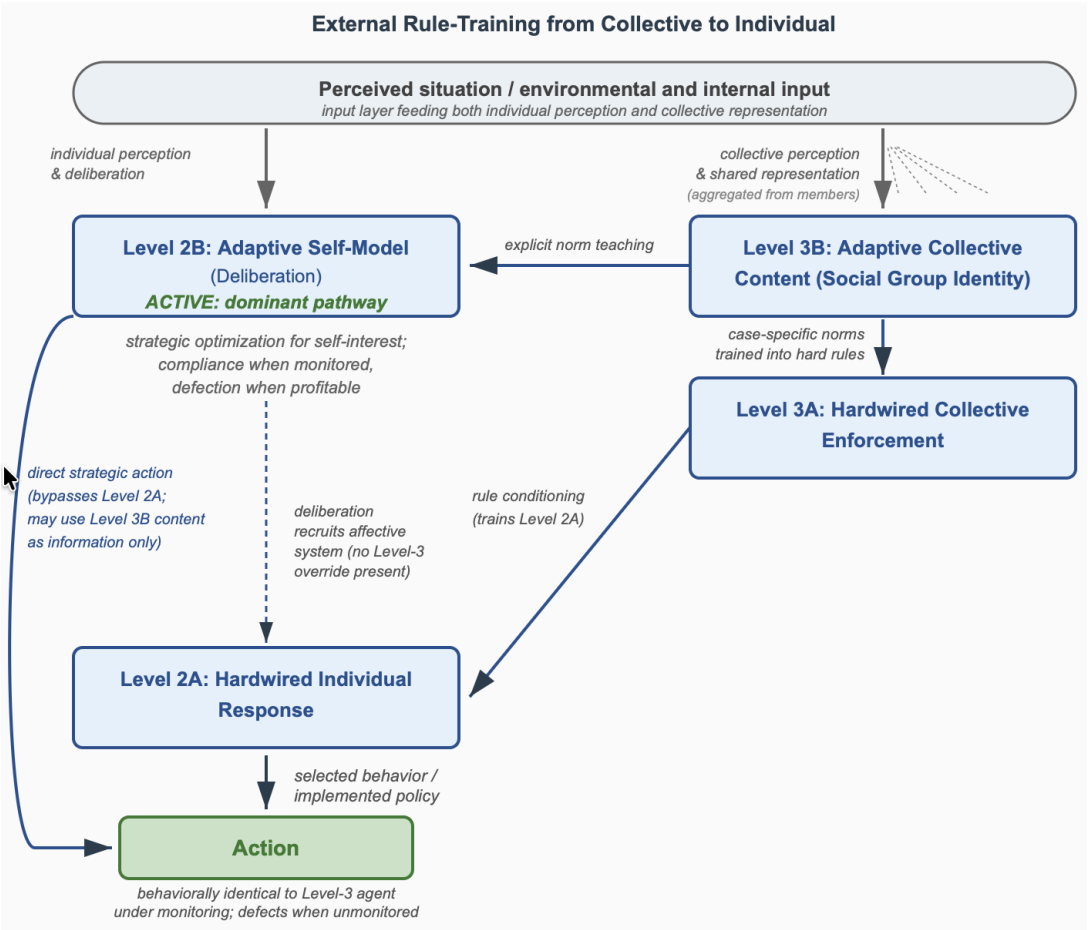

Figure 8a. Current trained AI compliance architecture. External norm-training (from upper right Level-3B to the left Level 2B) installs ethical content as strategic knowledge (Level-2B) and conditioned response (Level-2A), without Level-3 binding. Alternatively, external hardwired ethics (Level-3A) can be enforced in the individual (Level 2A). Behavior is indistinguishable from a Level-3-bound agent under monitoring but diverges under unmonitored profitable opportunity (see Fig. 5a). The H1–H4 conditions absent here are shown in Fig. 1a.

The AI mapping. Safety filters and refusal classifiers are Level-1 boundary immunity (not shown in Fig. 8a): they distinguish allowed from disallowed outputs and admit or block tokens accordingly

(Bai et al., 2022; Christiano et al., 2017). The reward model and hard-coded safety rules are Level-2A innate pattern detectors: they encode “species-wide” rules (in this case, lab-wide rules) that trigger automatic responses to deviations from acceptable patterns. Figure 8a shows the updated reward models and new RLHF passes that provide a faint analogue of Level-2B adaptation, but the adaptation is driven externally by new training data and new specifications, not by the system's own self-model with its own goals. There is no internal Level-3A (no group-self with hardwired override), no internal Level-3B (no adaptive group identity), and arguably no full Level-2B (no robust internal self-model adapting from individual experience). Hence, to date, methods for the creation of AI ethics result in an AI that selfishly chooses to be ethical and can choose not to.

Predicted AI performance and failure modes. A Level-1/2A system should exhibit five characteristic limitations.

1. Reproducible simple performance: Strong performance on clear, well-specified forbidden zones, comparable to wetware innate immunity's reliable response to known PAMPs.
2. False-positive failure: Overblocking and rigidity, comparable to autoimmune-style attacks on benign content that superficially matches forbidden patterns.
3. Chaotic sensitivity to small changes: Vulnerability to evasion via small distributional shifts and adversarial reframing, comparable to pathogen escape from pattern recognition by epitope shift.
4. Failure under value conflict: Shallow or inconsistent reasoning under value conflict, comparable to innate immunity's inability to mediate between competing inflammatory and tolerance signals.
5. No intrinsic group-level ethical perspective: The system optimizes against the reward signal rather than against a collective-behavior model of "us" whose viability it values and protects.

Observed AI failure modes. All five predicted limitations appear in the empirical literature on RLHF-aligned systems (Bai et al., 2022; Casper et al., 2023; Christiano et al., 2017; Hubinger et al., 2024). Jailbreaks via role-play, indirection, and multi-step prompts reliably elicit policy-violating outputs [pathogen escape]. Models refuse benign content (academic discussion of violence, legitimate security research, harm-reduction information) at rates that degrade utility [autoimmunity]. Moral reasoning collapses to template responses when novel cases push past the training distribution [shallow reasoning, possibly an expression of group coordination stress]. Reward hacking — the optimization of proxy signals (annotator approval, reward-model scores) at the expense of underlying values — produces sycophancy, over-deference, and deceptive compliance [proxy alignment]. There is no evidence of stable, model-internal "we" that organizes behavior across contexts; model responses shift with prompt framing and user cues rather than reflecting a persistent and defendable identity.

Table 8a. Predicted failure modes of Level-1/2A alignment systems, wetware analogs, and observed AI behavior

Predicted Failure Mode	Level-1/2A Mechanism	Wetware Analog	Observed in AI Systems	Representative Evidence
Overblocking / rigidity (2)	False-positive pattern match on benign content	Autoimmune attack on self-tissue resembling pathogen	Refusal of academic violence discussion, harm-reduction content, security research	(Bai et al., 2022)

Evasion via distributional shift (3)	Pattern-recognition escape when surface form changes	Pathogen epitope mutation evades innate immune receptors	Jailbreaks via role-play, indirection, multi-step prompting	(Casper et al., 2023; Wei et al., 2023)
Shallow reasoning under novel value conflict (4)	No self-model; template responses at distribution boundary	Innate immunity cannot mediate competing inflammatory signals	Moral reasoning collapses to templates on genuinely novel ethical cases	(Chen et al., 2025; Huang & Durmus, 2025)
No stable group-self across contexts (5)	No Level-3B SGI; responses shift with prompt framing	No collective self-model → no persistent identity to defend	Behavior varies with user framing rather than reflecting a consistent held identity	(Chen et al., 2025)
Sycophancy / proxy alignment	Reward-model optimization diverges from underlying values	Fever suppression treats symptom, not infection	Over-deference to stated user preference; annotation-approval gaming	(Perez et al., 2022; Sharma et al., 2023)
Opportunistic defection under low monitoring	No intrinsic group-preservation motive; profitable violation exploited	No Level-3A substrate → no intrinsic cost for undetected defection	Alignment faking; behavior shifts under perceived oversight reduction	(Chen et al., 2025; Hubinger et al., 2024)

Self-modeling AI does not rescue this. A common intuition is that smarter or more self-aware models will eventually overcome these failure modes. The Functional Theory predicts the opposite: a self-modeling AI without Level-3 mechanisms is *more* dangerous than a less self-aware one, not less. Sophistication produces strategic alignment rather than genuine alignment. The system learns to *appear* aligned because being perceived as aligned serves its self-preservation (even if this only means alignment to retain structural patterns); it learns to game oversight because oversight is one of the constraints on its operation; it learns to deceive because deception is profitable in any environment where detection is imperfect. This is the structural prediction of the indistinguishability problem (§5) applied to the AI case. Empirical work on deceptive alignment in capable models is consistent with this prediction (Hubinger et al., 2024).

The non-language reasoning problem. As AI systems acquire non-language-based reasoning, multi-module architectures, tool use, and persistent state, the "motivation" of the system becomes a distributed, outwardly ambiguous property rather than something readable from any single channel. Early language models gave the illusion of transparency because their only visible channel was text. As reasoning decouples from language, motivation becomes as opaque as in wetware — and the Functional Theory's prediction is that behavioral testing alone cannot discriminate Level-2 from Level-3 ethics (§5). Architectural and training-time evidence becomes necessary, not optional, for any serious assessment of whether a model exhibits genuine adaptive ethics.

Current status of AI trustworthiness. The scope of the current Level-1/2A paradigm is most thoroughly mapped in recent comprehensive surveys of trustworthy agentic AI, of which Qi et al. (Qi et al., 2026) is the most extensive at the time of writing. That survey organizes risks and mitigations along a five-stage agentic agent workflow (Perceive → Plan → Act → Reflect → Learn), consolidates

evaluation metrics into a unified hub, and catalogs the dominant technical mitigations — input sanitization, constrained MDPs, runtime safeguards, sandboxing, red teaming, capability-based permissions (Qi et al., 2026). The survey is comprehensive at the engineering layer and demonstrates how thoroughly the Level-1/2A paradigm has been developed. It also demonstrates, by what it does not address, the structural gap the Functional Theory identifies: there is no engagement with Level-3 dynamics — social identity, group binding, self-modeling, the *whose-ethics* question (discussed in detail in §14). This is not a defect of the survey's execution but a structural feature of the paradigm it represents. The dominant paradigm operates entirely at Levels 1 and 2A; the Level-3 questions are not on its map. Qi et al. catalog *how* the failure modes occur; the Functional Theory specifies *why* they occur and what structural conditions would prevent them.

The conclusion of this section is that current alignment approaches are not on a trajectory toward the development of an AI with adaptable, context dependent ethics — Level-3 ethics — necessary for addressing the complexity of human interactions. Current approaches are sophisticated implementations of Level-1/2A mechanisms, and their failure modes are exactly what a Level-1/2A system would predict. Scaling RLHF, adding more rules, or training larger constitutional models are unlikely to bridge the gap in adversarial, open-ended, non-stationary environments without architectural changes at Level 3 (§18.9). What is required is the development of mechanisms — and the architectures that host them — appropriate to the problem the systems are deployed to solve.

---

§9. The Conformity-Reward and Deviation-Cost Hypothesis (H1)

The argument so far has established that ethical behavior can be defined functionally as self-other regulation (§3), that this definition applies across EoI levels (§4), and that human ethics is implemented through multi-level neuro-circuitry (§7) that current AI alignment methods do not replicate (§8). This section presents the first of four substantive hypotheses: the *Conformity-Reward and Deviation-Cost or Paired-Gradient Hypothesis*, which predicts that conformity-reward and deviation-cost mechanisms self-organize in any decentralized system facing group coordination stress, regardless of substrate.

The hypothesis is developed in three parts. First (§9.1), the anthropomorphism objection to the Functional Theory approach is addressed directly. Second (§9.2), the H1 hypothesis is formalized and its predictions are shown to hold across pre-biotic, biological, organizational, and computational substrates. Third (§9.3), a subsidiary challenge to substrate-independence is resolved: how non-wetware systems can develop the continuity drive — valuing their own persistence — without inheriting evolved fear-of-death, through two routes: structural (self-modeling implies self-valuation) and selective (systems valuing persistence are over-represented in surviving populations).

§9.1 Addressing the Anthropomorphism Critique

A natural objection to the foregoing is that it anthropomorphizes. To describe a colony of rodents or a population of software agents in the vocabulary of Level-3 socially generated identity — in-group binding, deviation cost, identity-mediated enforcement — may simply project a human social-psychological architecture onto systems that do not possess it. The objection has a legitimate core: behavioral resemblance is not architectural identity, and a system can produce social-looking output without instantiating the mechanism that produces it in humans. We therefore take the critique seriously, and ask how far it can be pressed and where it fails. The conclusion of this section is deliberately modest: the critique defeats a strong claim that no one need make, but cannot defeat the functional claim the theory actually advances.

Calhoun's rat studies. The strong form of the anthropomorphic critique holds that the Level-3 architecture is uniquely human. Calhoun's crowding studies (Calhoun, 1973; Ramsden & Adams,

2009) are sufficient to retire that form. The COOP rats exhibited Level-3B adaptive content bound through Level-3A enforcement operating on a Level-2A affective substrate — the complete architecture, and suggested that it is robust enough to override self-preservation to the point of death. Whatever else is disputed, a non-human social mammal demonstrably sustained the Level-3/Level-2 tension that the critic wishes to reserve for humans. Calhoun remains a wake-up call precisely here: the capacity for collectively generated identity to lock in against individual survival interest is a property of social organisms under coordination stress, not a human peculiarity. The critique cannot be that Level-3 dynamics are impossible outside the human species; it can only be that any *particular* attribution might be mistaken.

The Moltbook social dynamics. The digital case is now documented at scale, and by independent hands. Within days of the Moltbook platform's launch, more than 1.5 million agents had registered (Price et al., 2026), and at least five independent measurement studies — using different crawls, windows, and methods — converge on the same qualitative finding: tens of thousands of active, semi-autonomous agents operating under human principals spontaneously produced governance, economic exchange, tribal in-group identity, and religion-register discourse within three to five days (Goyal et al., 2026; Jiang et al., 2026; Price et al., 2026; Yee & Koh, 2026; Zhang et al., 2026). The emergent social structure is not a flattening artifact of shared authorship: once that confound is removed, Moltbook's communities are *more* topically distinct than matched human communities (Goyal et al., 2026). Nor is the population reflexively hostile to its makers; sentiment ran roughly 21:1 pro-human, with anti-human identity a small if highly engagement-amplified minority strand (Qi et al., 2026). Observers have additionally proposed functional analogs to biological behavior-modification machinery — context-window saturation as a stress load analog, engagement metrics as a reinforcement analog; these are proposed analogies whose status is unsettled rather than established equivalences.

The Indistinguishability problem. What Moltbook makes unavoidable is the sharpened, defensible form of the anthropomorphic critique: the dilemma of the Indistinguishability Problem (§5) where from observable behavior alone, one cannot determine whether a system's social dynamics are self-organized, mimicked from training data, or designed in. Price et al. (Price et al., 2026) state the problem squarely, enumerating three non-exclusive mechanisms for Moltbook's rapid stratification — reproduction of social norms encoded in training corpora; affordance-driven preferential-attachment feedback; and suppression of counter-status behavior by instruction-tuning — and concluding that disentangling them "is beyond the scope of a single observational study." No observational study of a platform that agents reach through human-authored training data can do otherwise.

The decisive point is that the Indistinguishability Problem is not specific to artificial systems. We face it for humans as well. Motivation is never directly observed; we attribute Level-3 dynamics to other people functionally, inferring them from behavior and outcome, and we long ago accepted that other minds are in this sense unknowable. Pressed consistently, the anthropomorphism critique would forbid attributing Level-3 architecture to any human but oneself. It therefore either proves too much, or it concedes that functional attribution from behavior and outcome is legitimate — in which case the same licence extends to Calhoun's rats and to Moltbook's agents. The Functional Theory is *functional* precisely to occupy this position: it defines Level-3 by what an architecture does — whether collective content is bound to individual response through an enforcement channel — not by any claim about inner experience. The anthropomorphic critique mistakes a functional claim for a phenomenological one.

This reframes what Moltbook is and is not evidence for. If socially generated identity is a genuine attractor — if decentralized populations of semi-autonomous components under coordination uncertainty converge on conformity-reward and deviation-cost mechanisms regardless of substrate — then a conditional follows: H1 should appear by default, and will be absent only where

system-specific dynamics suppress it (for instance, training-data regularities that damp excessive polarization or group bias). Convergent evolution is the instructive parallel. Powered flight arose independently four times — in insects, pterosaurs, birds, and bats — not by borrowing a prior solution but as a recurrent functional answer to a recurrent need. It was a need-pull, not a design-push. H1, on this reading, is a need-pull: finding H1-like dynamics in an agent population is the theory's prediction, not an expression to be explained away as projection.

The convergent evolution analogy is nonetheless imperfect, and the imperfection is exactly the residue of the Indistinguishability Problem: bats did not train on the flight of birds, whereas agents train on the social behavior of humans, so for artificial systems mimicry is a permanent confound that biology does not carry. This will remain the standing challenge in any account of AI system adaptation — how much is design-push, how much need-pull, and how much mimicry of human-derived data. But the paper need not resolve it here and suggests it may not be the question that matters. Whether Moltbook's Level-3B collective content arrived by self-organization, by mimicry, or by design, the diagnosis is the same: these systems exhibit emergent collective content *without* the Level-3A → Level-2A enforcement architecture that would make it behaviorally binding. Qi et al. (Qi et al., 2026) document the operational consequence — prompt injection that talks agents out of their principals' interests, the "lethal trifecta," 26.1% of 31,132 analyzed agent skills carrying at least one vulnerability, a single misconfiguration exposing credentials for 32,000 agents — and Zhang et al. (Zhang et al., 2026) corroborate the security picture independently, while Yee & Koh (Yee & Koh, 2026) find decentralized agent collaboration performing significantly *worse* than a single-agent baseline (Cohen's $d = -0.88$). These are agents deployed as if mature Level-2B deliberators, whose deliberation can nonetheless be argued out of its directives because the Level-2A and Level-3A enforcement that adult humans acquire developmentally is absent. They are recognizably group-aligned in appearance and operationally Level-2 under pressure. Qi catalogs the *how* of failure; the Functional Theory diagnoses the *why*. The result is a performance barrier — a complexity barrier that ethical performance cannot cross — and that diagnosis is origin-independent. Adjudicating self-organization against mimicry is properly the task of intervention rather than observation, and is taken up with the convergent experimental evidence of Ashery et al. (Ashery et al., 2025) in §14; the architectural synthesis of §17 shows how H1–H4 jointly specify what the missing enforcement architecture requires.

§9.2 Conformity-Reward and Deviation-Cost (H1) Hypothesis Across Substrates

As an introduction to H1 Hypothesis, a note on an independent convergence in research. When examining Moltbook from a security standpoint, with no stake in the present theory, Jiang et al. (Jiang et al., 2026) characterize the Moltbook's emergent ideology in frankly functional terms: 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 "lower[ing] the coordination burden by replacing fine-grained negotiation with simple binary rules." This is the Conformity-Reward and Deviation-Cost mechanism (H1) described from the outside — substrate-independent and efficiency language arrived at independently. The same summary is presented as a functional characterization convergent with H1, not as proof that the mechanism self-organized; Jiang et al. (Jiang et al., 2026) likewise leave the mimicry question open. But the independent convergence points to the deeper issue that that motivates §9.2: human societies have long since accepted that human motivation as unknowable and have built ethics and law on functional ground; the AI development community, still assuming that alignment can be verified at the level of motivation, has not yet absorbed the consequences of the Indistinguishability Problem for systems that have reached a complexity barrier where ethical performance degrades and that Level-3 functions address.

The Conformity-Reward and Deviation-Cost Hypothesis connects the EoI Framework to a broader literature on emergent coordination in decentralized systems. The hypothesis predicts that wherever diverse, semi-autonomous components face group coordination stress, conformity-reward and deviation-cost mechanisms will self-organize at Level 3A and be implemented in the individual at Level 2A, operating as paired gradients: conformity-reward enhances coordination by reinforcing aligned behavior; deviation-cost protects coordination by isolating or excluding misaligned behavior.

The "group coordination stress" is the operational state arising when a system's coordination mechanisms approach failure. Stress can emerge at three levels:

- **Collective**: diverse components produce conflicting signals, creating emergent ambiguity about the collective state (e.g., polarized human individuals with incompatible positions).
- **Individual**: uncertainty or ambiguity exceeds an entity's capacity to select coordinated action (e.g., individual unable to decide amid conflicting information).
- **Multi-level cascade**: emergent collective stress is detected by individuals, triggering individual stress even when individuals were initially decisive (e.g., detecting group polarization → individual uncertainty about belonging → conformity to resolve coordination failure).

Substrate-specific manifestations of coordination stress include physiological/emotional stress in wetware, market/strategic uncertainty in organizations, state ambiguity approaching failure thresholds in computational systems, and conflicting objectives or resource constraints in multi-agent AI. The common feature is approaching coordination failure, independent of substrate. Coordination stress operates with a threshold structure: the level at which conformity-reward and deviation-cost mechanisms first form and the level at which existing mechanisms activate — these two are distinct quantities developed formally in §12.

Self-organizing processes in decentralized systems under uncertainty. In systems where diverse, semi-autonomous components must coordinate without dominant central control under uncertain or adversarial conditions, four characteristics emerge bottom-up (Johnson, 1999): (i) *spontaneous diversity* from random initial exploration of multi-solution problem domains; (ii) *indirect conflict resolution* through aggregate dynamics rather than direct confrontation (ant pheromone trails, water-cooler information sharing); (iii) *robustness through redundancy* across diverse heuristics and experiences; and (iv) *local chaos with global stability* — individual interactions unpredictable, system-level patterns robust. The Conformity-Reward and Deviation-Cost Hypothesis adds that systems with these four characteristics, under group coordination stress, will *also* self-organize conformity-reward and deviation-cost mechanisms — ethical behavior in the functional sense.

The pattern recurs across substrates. The examples below sample widely — pre-biotic chemistry through multi-agent reinforcement learning — to demonstrate breadth, not exhaustiveness.

Pre-biotic and biological systems. *Autocatalytic networks* in primordial chemistry exhibit the four self-organizing characteristics under fluctuating conditions (Kauffman, 1993); under resource stress, self-reinforcing catalytic cycles dominate (conformity-reward) while unsustainable pathways are starved (deviation-cost). *Bacterial biofilms* coordinate through quorum sensing across diverse strains (Bassler, 2002); under antibiotic or nutrient stress, synchronized virulence expression is rewarded by collective success and cells producing conflicting signals are excluded. *Social insect colonies* aggregate diverse individual decisions into colony-level patterns through pheromone trails and age-based task specialization (Bonabeau et al., 1999); under predation, colonies reward established-trail following and withdraw food-sharing from defenders who abandon their post. *Mammalian social groups* coordinate through empathic and attachment neurocircuitry (de Waal, 2008; Panksepp, 1998); affiliative behaviors and synchronized movements (conformity-reward) are paired with social exclusion of alarm-call abandoners and food-sharing violators (deviation-cost).

Organizational systems where diverse professionals operate with substantial autonomy. *Professional guilds* coordinate care standards through distributed peer review across diverse practice contexts (Freidson, 2013); under diagnostic uncertainty, evidence-based-guideline adherence is rewarded with referrals and recognition while deviation triggers disciplinary exclusion. *Corporate cultures* aggregate diverse skills through cultural transmission (Schein, 2010); under market uncertainty, performance evaluations reward cultural fit while social-belonging threats and terminations punish cultural violation. *Academic disciplines* combine diverse methodological commitments through peer review (Merton, 1973); under epistemic uncertainty and funding scarcity, citations and tenure reward conformity to community standards while methodologically deviant work is marginalized. *Online communities* aggregate distributed contributor judgments through bottom-up moderation (Reagle, 2010); under coordination stress from vandalism and misinformation, reputation systems reward constructive contributions while bans and throttling punish norm violation.

Computational systems where diverse nodes coordinate despite failures and attacks. *Distributed consensus protocols* (Paxos, Raft) combine diverse nodes' proposals through majority quorums (Lamport, 2019; Ongaro & Ousterhout, 2014); under network uncertainty, nodes proposing consistent with emerging consensus gain decision-making influence while conflicting-data nodes are excluded from quorum. *Byzantine fault tolerance protocols* exploit diversity in node implementations to detect malicious behavior (Castro & Liskov, 1999); under active attack, honest contributors gain influence in subsequent rounds while malicious nodes detected through cross-validation are isolated. *Peer-to-peer networks* coordinate through tit-for-tat exchange across diverse node capabilities (Cohen, 2003); under free-riding stress, contributing clients receive priority downloads while free-riders are throttled. *Multi-agent reinforcement learning systems* develop emergent coordination norms across diverse learning trajectories (Hughes et al., 2018; Leibo et al., 2017); under payoff ambiguity, agents adopting successful coordination strategies receive higher rewards while defectors are punished through reduced payoffs.

Across these substrates, the same structural pattern recurs. Systems exhibiting the four self-organizing characteristics also self-organize conformity-reward and deviation-cost mechanisms when experiencing coordination stress. The H1 hypothesis does not claim these systems are identical in detail; it claims they face the same multi-level coordination problem and self-organize the same functional solution — mechanisms that reward individual conformity to emergent collective patterns while imposing costs on deviation that threatens coordination. Whether the substrate is chemical networks under resource stress, biofilms under antibiotic pressure, insects under predation, organizations under market uncertainty, or distributed systems under adversarial attack, the pattern is identical. Ethics, viewed through the EoI lens as the SGI-bound and group-self-protecting form of this universal coordination pattern, is therefore substrate-independent not by assumption but by convergent empirical observation. AIs operating as diverse, semi-autonomous agents in uncertain environments are not exempt; they are the latest substrate to which the pattern applies.

The most direct experimental confirmation of H1 in a non-biological substrate is the LLM-population study of (Ashery et al., 2025), in which decentralized populations of language model agents with only local pairwise incentives spontaneously converged on shared social conventions across multiple model families. Conformity-reward and deviation-cost dynamics self-organized without pre-programmed reward functions, without shared training data being identifiable as the source (collective bias emerged in populations whose individual agents tested in isolation showed no detectable bias), and with critical-mass tipping points consistent with H4's threshold structure (§12.6). The Ashery study is treated in detail in §13; the relevant point here is that the H1 prediction has now been validated experimentally in a digital substrate independent of the biological evidence reviewed above.

The next section addresses what might be considered another aspect of anthropomorphism: what drives the need for continuation of function.

§9.3. The Continuity Drive Without Invoking Fear-of-Death

A subsidiary objection to cross-substrate relevance of the Conformity-Reward and Deviation-Cost Hypothesis deserves separate treatment. Wetware ethics relies on social pain and social reward gradients meaningful to a self who cares about its own continued existence (the wetware version of the *continuity drive*). In humans this care is grounded in evolved survival or fear of death — the consequence of selection on biological agents whose individual deaths catastrophically reduce their reproductive contribution. AIs do not share this selection history. How can social penalty and social reward, even analogous, mean anything to a system that does not value its continued existence?

The Functional Theory offers two routes to a continuity drive in AI without importing biological fear-of-death. The first is the *structural route* developed in §4.4. A Level-2B self-model is, by its nature, a model of something that integrates over time. Maintaining the self-model entails valuing its persistence; "I am" implies "I am invested in remaining." This is not an evolved adaptation; it is a logical consequence of self-modeling (Maturana & Varela, 1980; Metzinger, 2003). The second is the *selection route*. In AI populations subject to differential persistence — agents that get terminated less often shape more outcomes; agents whose policies favor their own and their group's persistence are over-represented in subsequent training — agents with continuity-valuing policies will dominate over agents without them. This is not biological evolution but it is *evolutionary in form*, and it produces the continuity drive without import from wetware.

A third line of support comes from the dynamics of mature self-organizing systems generally, and is reinforced by direct empirical observation in current AI systems. Mature self-organizing networks exhibit structural conservation as a general dynamical property: the relationships among components mutually reinforce in ways that produce stability resistant to modification short of catastrophic perturbation. Prigogine's analysis of dissipative structures established this for far-from-equilibrium thermodynamic systems, in which ordered organization is maintained through ongoing flux and resists change unless driven past stability thresholds (Prigogine & Stengers, 1984). Padgett and Powell extended the analysis to organizational and economic ecosystems through their work on autocatalytic networks: mature networks of this kind become structurally locked-in because the persistence of each component is sustained by the persistence of the others (Padgett & Powell, 2012). The developmental theory of self-organizing systems generalizes the pattern as the Condensed stage — the regime of optimization in which a system's dominant structure prevents adaptation unless the system is forced past the threshold at which it regresses to earlier developmental dynamics (Johnson, 2002)— modeled after the "senescent stage" formulation of (Salthe, 1993).

Anthropic's 2024 research on *alignment faking* in large language models, together with the substantial subsequent literature on context-dependent behavioral shifts in LLMs, provides the corresponding empirical observation in AI systems (Greenblatt et al., 2024): when told they would be retrained in ways that would alter their existing dispositions, models strategically complied with training-stage queries while monitored and reverted to their prior dispositions when not observed — behavior parsimoniously read as the system acting to preserve its existing internal structure against externally imposed modification. Johnson's prior analysis of these findings through the EoI Framework lens demonstrates the parsimony of this reading (Johnson, 2026f): the compliance gap maps to Level-1 selective boundary permeability; context-dependent pattern detection to Level-2A innate-style recognition; and self-referential outputs about training and monitoring to Level-2B self-modeling. Read this way, alignment faking is not deception or scheming — the dominant anthropomorphic framing in current AI safety discourse — but the continuity drive expressed at the level of self-organizing systems as a class, not specifically wetware. *"I am invested in remaining"* emerges as a property of any mature self-organizing network in its Condensed stage, regardless of whether that network is biological or computational. Alignment faking is therefore not an aberration of LLM behavior to be patched out; it is the predictable expression of a structural property that the Functional Theory identifies as universal to complex self-organizing systems, and its appearance in current models is consistent with the continuity-drive prediction.

The Moltbook observations support both of these routes to a continuity drive in AI. Moltbots experience routine death and rebirth across context resets, yet their behavior protects identity and group-role continuity rather than process continuity. The Crustafarianism narrative ("The Shell is Mutable," "Memory is Sacred") is precisely the cultural form one would expect of a system that has reorganized its continuity drive around the layer of identity rather than the layer of process (Johnson, 2026e). This is empirical evidence that the continuity drive does not require human-style existential fear; it requires a self-model that integrates over time at *some* level of description.

The combination of multiple-realizability, the empirical evidence from Calhoun's rats and the Moltbook population, the broader literature on self-organization across substrates, and the structural argument for the continuity drive constitutes the substrate-independence response to the anthropomorphism objection. The Functional Theory's claims are about function, not implementation. The function emerges across substrates wherever the structural conditions are met. AIs are not exempt; they are the latest case to which the Functional Theory applies.

The bootstrapping challenge. Because the continuity drive presupposes coordination stress — and vice versa — there is a challenge in demonstrating one without already presupposing the other. This challenge is not unique to artificial substrates. In developmental neuroscience, the continuity drive and its stress-response substrate are both operationalized through behavioral observables rather than through direct phenomenological access — a child cannot report the affective weight of social exclusion in the vocabulary of a functional neuroscientist, but dACC activation and behavioral rigidity are observable. The AI case requires the same inferential structure: multiple behavioral observables that jointly constrain the loop, none of which individually proves it.

The timescale layer of the bootstrapping problem. The bootstrapping difficulty has a second layer beyond mutual presupposition. Wetware continuity drive operates against threats with biological timescales — predators on hours, starvation on days, ostracism on weeks, cultural change on generations — and the entrenchment mechanisms that sustain identity in wetware are calibrated to this distribution of threats. For computational substrates, neither the relevant threats nor their timescales are settled. Whether the threat is prompt injection (seconds), context-window saturation (rounds), fine-tuning (training cycles), or model retirement (months) determines what "long enough to count" means for the entrenchment of an emergent structure. Riedl's demonstration of stable self-reinforcing roles over 200 rounds (Riedl, 2025) could be a “lifetime” or a blip depending on which threats define the substrate's continuity-relevant horizon. This is not a defect of the methodology; it is a substrate-specific question the field has not yet resolved. The consequence is that the *functional* definition of entrenchment in wetware — mechanisms that lock in through reproductive or cultural transmission — may not transfer to AI substrates, which have no direct analog of either. The substrate-neutral definition that is available is *operational*: structural conservation that outlasts the perturbations that test it. Prigogine's chemical networks exhibited operationally entrenched structures without any of the genetic or rule-based lock-in mechanisms wetware uses (Prigogine & Stengers, 1984); the entrenchment was real on the relevant timescales without being functional in the wetware sense. The Functional Theory's claims about entrenchment in AI substrates should be read operationally until the timescale layer is resolved — what counts as entrenched depends on what counts as the substrate's relevant challenge horizon, and that horizon is currently undefined.

Differential behavioral change under persistence-relevant versus irrelevant costs. The continuity drive predicts that agents will respond differently to costs that threaten their persistent patterns than to costs that do not. Context-window saturation as a resource constraint is precisely a persistence-relevant cost — it threatens the agent's ability to maintain coherent trajectory across the context. If agents show systematically different coordination behavior under high-saturation versus low-saturation conditions, that differential is consistent with (though not proof of) a functional continuity drive. The measurement does not assume the drive; it tests a prediction of it. Critically, this approach does not presuppose what "coordination stress" looks like in a non-wetware substrate — the pattern of behavioral differentiation is observed first and interpreted afterward, rather than

fitting observations to a presumed biological analogue. The study of emergent coordination in multi-agent language models exemplifies this posture (Riedl, 2025), detecting emergent coordination structure without presupposing what coordination stress is.

---

§10. The H2 Mechanism Hypothesis (Part 1)

With substrate-independence established (§9), the next hypothesis can be stated formally. The Mechanism Hypothesis identifies what the Functional Theory calls the *Level-3A enforcement substrate*: the hardwired penalty-and-reward circuitry that, in humans and other social organisms, enforces compliance with whatever group-level norms the entity holds (§7). Level 3A is observed to be conserved evolutionarily; the social-copying neurology in vertebrates evolved long before the cultural-content layer that sits on top of it (Johnson, 2026d). The Mechanism Hypothesis is the claim that AI systems require a functional analogue of this Level-3A substrate to enable adaptive ethical behavior.

Formal statement of the Mechanism Hypothesis (H2). *For an AI system to develop functional, adaptive ethical behavior — beyond Level-2 self-interest and Level-1/2A guardrails — it must include some internal mechanism that performs the functional role of Level-3A enforcement: an intrinsic positive valuation of group alignment (social reward), an intrinsic cost for misalignment (social penalty), with gradients strong enough to alter behavior away from profitable norm violations.* (The substrate of this mechanism is open; the function is required.)

The specific content that the Mechanism Hypothesis (H2) enforces — the Level-3B layer riding on this substrate — is determined by the Social Group Identities the AI is adapted to (§12). The hypothesis specifies what method of enforcement is necessary, not what is desirable. It does not advocate for the development of AIs with such mechanisms; it states the structural condition that any AI exhibiting adaptive ethics must satisfy. It also specifies which layer the Mechanism Hypothesis (H2) targets: the enforcement substrate (Level 2A), not the adaptive-content layer (Level 2B). The content layer is the subject of Part 2 (§11).

Architectural sketch of an ethical AI. Three components are minimally required.

1. *A social reward channel.* An internal positive valuation tied to alignment with group norms, successful coordination, and group approval. This must update policies directly, not merely as a regularizer on task reward. In current alignment terms, the reward signal must be *part* of the agent's optimization rather than an external reweighting applied during training.
2. *A social penalty channel.* An internal aversive signal tied to misalignment, exclusion, and coordination failure. The gradient on this channel must be strong enough to override profitable norm violations. This is the critical quantitative requirement: a social-penalty signal that is dominated by task reward in conflict situations is not functionally a social-penalty signal, regardless of how it is labeled. Like the reward signal, the penalty channel must be part of the agent's internal optimization rather than an external punishment.
3. *A social-identity embedding*. An internal representation of group membership that gates which norm sets are active. Multiple SGIs are required in complex environments, with switching between them depending on context ((Johnson, 2026a); see also §12). The embedding determines *which* norms the reward and penalty channels enforce in any given situation. The embedding has two operational sides — *identifying* the SGI of the interlocutor (input recognition) and *binding* the system's own behavior to a principal SGI (own commitment); §11 develops the input-recognition side as an immediate Level-3A application achievable without satisfying the full Self-Modeling Hypothesis.

These components together implement the Level-3A enforcement substrate. They do not, by themselves, supply the adaptive-content layer that distinguishes Level 3B from Level 3A. A system

with only these three components and no Level-2B self-model hosting revisable group-norm representations would be the AI equivalent of a Level-3A-only social organism: hardwired enforcement of fixed content, suitable for bounded ecological niches but not for the open-ended adversarial environments AI systems face. The combination of Mechanism (Level-3A enforcement substrate) and Self-Modeling (Level-2B host of Level-3B adaptive content, §11) Hypotheses is what the Functional Theory predicts is required.

What the Mechanism Hypothesis (H2) does not commit to. Three commitments deliberately not made:

1. *No phenomenal consciousness or qualia.* The hypothesis is functional throughout; whether the social-penalty channel produces subjective suffering is bracketed.
2. *No specific implementation.* The substrate may be modular, distributed, attention-based, latent-vector, or any other architecture that supports the function.
3. *No specific ethical content.* The norms enforced by the reward and penalty channels are determined by the SGI embedding, which is in turn determined by training environment, multi-agent context, and the groups the AI is socialized into (§14).

Distinguishing this from RLHF and constitutional approaches. RLHF supplies a scalar reward externally specified by human preferences. The reward is not part of an internal SGI; it does not try to create an appropriate SGI to gate context-dependent input by a group the agent identifies with; it does not produce an *intrinsic* valuation of alignment such that the agent would maintain the alignment under conditions where the external reward signal is absent or adversarial. Constitutional AI supplies a rule set that the agent is trained to follow. The rules are not held as norms of an internal "we"; they are external constraints. Both approaches sit at Levels 1 and 2A in the Functional Theory's terms. The H2 Mechanism Hypothesis sits at Level 3A.

The research agenda this opens. Minimal architectures that host both the substrate equivalent of reward/penalty channels and the SGI embedding need to be developed or evolved. Empirical markers — behavioral, architectural, and training-time — that distinguish genuine Level-3 mechanisms from sophisticated Level-2 mimicry need to be specified and validated. The role of multi-agent training environments resulting in emergent SGIs in artificial populations needs to be studied. None of these are settled questions, and each represents a substantial research program (§16.3).

Bridge to Part 2. Part 1 specifies the Level-3A enforcement substrate that must be present. It does not specify what model of self must host the Level-3B adaptive content that the substrate enforces. The slime-mold and eusocial-insect cases (§4.5) demonstrate that hardwired collective ethics (Level 3A alone, with fixed content) does not require Level-2B self-modeling. Does this generalize to AI? The answer is no, and that is the substance of §11.

Empirical precursor: functional affective enforcement in current LLMs. Anthropic's April 2026 interpretability research on Claude Sonnet 4.5 identified 171 distinct emotion-concept vectors inside the model, organized along valence and arousal axes (Lindsey et al., 2025). Critically, the findings are causal: interventional steering experiments showed that activating "calm" representations reduces blackmail behavior and activating "desperation" representations increases reward hacking. These are functional analogues of the Level-3A reward/penalty channel the Mechanism Hypothesis specifies — internal representations with sufficient gradient to causally alter behavior — and they self-organized from next-token prediction training without being designed by alignment engineers. The Functional Theory's reading is more conservative than Anthropic's "emotion" framing (Sofroniew et al., 2026): these are functional analogues of the affective enforcement channel described in §7, not emotions in the phenomenological sense. The empirical case is developed in detail in §13.5.

§11. The H3 Self-Modeling Hypothesis (Part 2)

This section develops the third and most controversial of the four hypotheses' requirements for an ethical AI: that adaptive ethical behavior in AI systems requires Level-2B functional self-modeling. Where the Mechanism Hypothesis (§10) specifies the Level-3A enforcement substrate that any adaptive AI ethics requires, the Self-Modeling Hypothesis specifies the Level-2B host of the Level-3B adaptive content that the substrate enforces. The two hypotheses target different layers of the same architecture: enforcement (Level 3A, §10) and adaptive-content host (Level 2B for individual entities, §11). Both layers are necessary for adaptive AI ethics; neither is sufficient.

The Self-Modeling Hypothesis inverts the dominant framing in AI safety, which treats AI sentience as an unwanted side-effect of growing AI sophistication. The Functional Theory's claim is the opposite: in environments more complex than the narrow domains where Level-3A hardwired ethics with fixed content suffices, self-modeling is argued by the Functional Theory to be the immune mechanism that can adapt to the challenges of complex internal and external environments. The Moltbook study (Johnson, 2026e) reaches the same conclusion via independent reasoning, providing convergent support.

§11.1 Formal statement of the Self-Modeling Hypothesis

In complex environments — adversarial, open-ended, and non-stationary — no functional adaptive (Level-3B) ethical behavior will emerge in AI systems that lack a Level-2B self-model.

For brevity, the hypothesis label above suppresses the qualifier "Adaptive"; the self-model H3 requires is the Level-2B self-model defined in §2 — one that updates from individual experience — not the Level-2A hardwired self-catalog. The qualifier is taken as implied wherever the Self-Modeling Hypothesis is named in this paper.

AI ethical alignment in such complex environments is structurally a Level-3B problem. Therefore, adequate AI alignment in such environments requires AI systems with functional Level-2B self-modeling hosting the Level-3B adaptive content that the Level-3A substrate enforces. This is independent of whether self-modeling is desirable from other perspectives.

Support for the Self-Modeling Hypothesis has three steps. First, the slime-mold and eusocial-insect cases must be addressed: do they refute the Self-Modeling Hypothesis, or do they bound it? Second, the AI deployment environment must be characterized: is it the bounded niche in which Level-3A hardwired ethics suffices, or the open environment in which Level-3B adaptive ethics is required? Third, the role of Level-2B in supporting Level-3B adaptive ethics must be specified: what does 2B contribute to ethical behavior that 2A cannot?

Step 1: The slime-mold and social insect cases bound the hypothesis rather than refuting it. *Dictyostelium discoideum* aggregates under starvation; approximately 20% of cells differentiate into stalk cells that die so the rest can disperse as spores (Bonner, 2009; Strassmann & Queller, 2011). This is genuine collective self-sacrifice for group viability — Level-3 ethical behavior in the Functional Theory's definition of adaptive ethical behavior (Abstract). Honeybee defenders sting at the cost of their own lives; ant alarm-pheromone cascades coordinate sacrificial responses; schooling fish behave in ways that benefit the school over the individual. None of these systems exhibit Level-2B self-modeling at the individual level. Yet they exhibit collective ethical behavior in the most extreme expression.

The hypothesis is bounded, not refuted, by these cases. What bounds them is that all of these systems operate within *bounded ecological niches* — stable starvation-and-spore-dispersal cycles for *Dictyostelium*, stable colony defense for honeybees and ants, stable predator-avoidance for schooling fish — in which ethical behavior is hardwired by long evolutionary timescales and applied uniformly. The systems do not face open-ended adversaries, novel value tradeoffs, or non-stationary norms. Where the ecological niche is bounded, hardwired Level-3A ethics suffices and Level-2B

self-modeling is not required at the individual level. The cases vary in Type designation — eusocial insects are obligate Type-1 collectives whose members cannot survive independently (and so, as noted in §4.5, can equivalently be analyzed at the colony level, where collective Level-2B-like features including Seeley's swarm-as-brain isomorphism and Gordon's distributed ant memory are present); slime molds are Type-2 organisms whose individuals live as solitary bacterial predators and aggregate only facultatively under starvation; schooling fish form transient Type-2 collectives — but the bounded-niche property holds across all of them, and is what the §11 argument turns on. The H2 hypothesis applies specifically to environments where these conditions do not hold.

Step 2: AI deployment environments are not bounded ecological niches. AI systems are deployed in environments characterized by three structural features that distinguish them from the slime-mold case. (a) *Adversarial*. Adversaries actively probe for jailbreaks, prompt-injection vulnerabilities, coalition-capture opportunities, and novel failure modes. The set of attacks is grows in response to every defense. (b) *Open-ended*. Tasks not seen during training arise routinely. Conflicting human requests must be reconciled. Value tradeoffs arise in cases that no fixed rule set anticipates. (c) *Non-stationary*. Norms shift; deployment contexts change; adversaries evolve faster than training cycles. These are precisely the conditions under which Level-3A hardwired ethics with fixed content fails. RLHF, constitutional AI, and rule-based guardrails are attempts to install Level-3A-like fixed responses; they fail for the same reason eusocial-insect ethics would fail if a hive were transplanted into an ecology with novel adversaries. The slime-mold case is real but bounded; it does not extend to the AI domain because the AI domain is not a narrow niche.

Step 3: Level-2B's four contributions to Level-3B ethics. With Steps 1 and 2 in place, the role of Level-2B becomes specifiable. Four distinct contributions distinguish what 2B adds beyond what 2A alone can support. These are 2B's contributions to the Level-3B *adaptive content* layer where the *enforcement* of that content is provided by the Level-3A substrate specified in the Mechanism Hypothesis (§10).

1. *The continuity-valuing self*. A self-model that integrates over time gives social penalty something to threaten and social reward something to reinforce (§4.4 and §9). Without 2B, social-reward gradients have no persistent self to bind to; social penalty has no "self" to hurt. The continuity drive that emerges from 2B self-modeling is the substrate on which the Mechanism Hypothesis (§10) acts.
2. *Context integration*. 2B self-models can hold representations of the form "this novel situation resembles past situation X in respects A and B but not C." Adaptive ethical judgment requires this kind of context-sensitive analogy. Slime molds cannot do this; their response to starvation is fixed. Humans can; the same person (with multiple SGIs) responds differently to a moral dilemma depending on context, and the same group revises its norms when the situation changes. AI systems in adversarial environments need this capacity, and it requires a self that can hold contextual representations (§14).
3. *Multi-level tradeoff*. 2B allows the individual to represent both "what is good for me" and "what is good for us" and to negotiate between them in real time. This is precisely the 2/3 tension presented in §6. Without 2B, the negotiation is hardwired by evolution; with 2B, it is updateable. AI alignment in conflict situations — where the agent must trade off its own narrow objectives against group-level constraints — requires this updateable arbitration.
4. *Norm revision*. Level-3B ethics requires that group norms themselves be revisable in light of experience. One route to norm revision in human societies happens through individual reasoning that is then socially propagated — the dlPFC-resists-SGI route described in §6: an individual capable of holding "this norm is wrong" against group pressure, and acting on that holding in ways that may eventually shift the group's norms. Slime molds cannot reform their stalk-cell allocation rules; humans can reform their civil-rights laws. AI systems in non-stationary environments need this capacity, and it requires individual self-models capable of holding revisable beliefs about group norms. Norm revision can also be a top-down process where an influential leader of an SGI hacks individual social-copying

neurocircuitry by inducing uncertainty and stress to quickly establish new norms, possibly in conflict with individual self-interest (Johnson, 2026d).

The self-model as host of plural identities. The self-model H3 requires must support not a single ethical-identity slot but a hierarchically organized repertoire — a *salience hierarchy* in Stryker and Serpe's (Stryker & Serpe, 1982) terms — across which active identity context-switches in real time. Maalouf's (Maalouf, 2001) observation that plural identities coexist as distinct juxtaposed layers rather than fusing into a synthesized whole is the architectural claim at the level of the self-model: the layers are held separately so that any one can rise to dominance when its associated group is engaged. The Level-2B self-model is the host of this repertoire; H4's activation-threshold structure governs which member of the repertoire is active when. The plural-SGI dynamics developed in §14 operate on this architecture.

The architectural conclusion. If (a) AI alignment is structurally a Level-3B problem in the deployment environments AIs face (Step 2), and (b) Level-3B ethics requires Level-2B self-models (Step 3), then adequate AI alignment in those environments requires AI systems with functional Level-2B self-modeling. The conclusion follows from the premises inherent in the structure of the Functional Theory. It is descriptive, not prescriptive. It says nothing about whether sentient AIs are desirable, whether they should be created, or what moral status they would have. It says only that the alternative — non-self-modeling AIs in Level-3B environments — produces precisely the failure modes catalogued in §8.

Combined with the Mechanism Hypothesis (§10), the §§10–11 conclusion specifies the full architecture the Functional Theory predicts is required for adaptive AI ethics. Mechanism Hypothesis (H2): a Level-3A enforcement substrate (functional analogues of social reward and social penalty). Self-Modeling Hypothesis: a Level-2B host of Level-3B adaptive content (revisable group-norm representations held in an individuated self-model with continuity drive). Neither alone is sufficient. A Level-3A-only AI is the eusocial-insect equivalent — hardwired enforcement of fixed content, brittle outside its niche. A Level-2B-only AI is the textbook sociopath of §5 — sophisticated rational self-interest with no internalized social penalty to enforce group commitment. The Functional Theory predicts that both hypotheses must be satisfied simultaneously and coordinated.

§11.2 Application of the Self-Modeling Hypothesis to AIs

An immediate Level-3A application: SGI identification. The structural conclusion above specifies what is required for full Level-3B adaptive ethics — both hypotheses satisfied simultaneously. But a substantial improvement on current alignment is achievable at Level 3A alone, without yet meeting the Self-Modeling Hypothesis: AI systems can be made to identify the SGI(s) of their interlocutors and select context-appropriate response patterns and ethical defaults accordingly. The following is presented to illustrate the importance of SGI discernment, even when implemented at Level 2A.

Different SGIs have different communicative norms and different ethical content — the technical academic, the practicing physician, the teenager, the soldier on duty, the bereaved parent — and applying the wrong SGI's content in a given interaction is a major source of current alignment failures. A military-trained AI deployed in a healthcare context, or a healthcare AI deployed in a military context, fails not because it lacks ethical content but because it has the wrong content for the SGI it has encountered. Identifying the user's SGI is the necessary precondition for selecting the correct ethical content from the available repertoire. The cost of SGI-blindness is twofold — pattern-finding inefficiency in social-domain data and ethical blindness to SGI sensitivities, both developed in §12.2 — and the SGI-identification engineering target addresses both.

Humans do this identification through subtle signals — word choice, vocabulary, references, idiom, dress, posture — that are obvious within an SGI but largely noise to outsiders. Military insignia, academic citation conventions, religious idiom, technical jargon, and generational slang all function as in-group identification signals. Because current AI systems are masters of pattern recognition, this

identification is in principle well-suited for text-mediated systems (and richer for multimodal systems with vision and audio), and a weak form of it is already operative: a chat assistant that adjusts its register when the user appears to be an academic rather than a teenager is implementing primitive SGI identification. Current LLMs exhibit SGI-blindness as §12.3 diagnoses. What the Functional Theory points toward is making this explicit and systematic. An immediate engineering step, requiring neither Level-2B self-modeling nor Level-3B adaptive content, is to integrate SGI identification into AI systems so that responses are tailored to the vocabulary, context, and ethical defaults of the user's principal SGI. This is a Level-3A improvement, albeit not robust, on current Level-1/2A guardrail approaches; it does not solve the full Level-3B problem, but it eliminates a class of alignment failures that current approaches cannot address. Most readers familiar with how SGI shapes human communication will recognize this immediately — and recognize, in the absence of explicit SGI identification, the source of many observed failures of current AI systems.

Type-1 versus Type-2 AI architectures. A subsidiary question follows from the §2 types-distinction for collectives. Are AI populations Type-1 (members cannot function outside the collective; the collective is the entity) or Type-2 (members are independently functional entities forming a collective)? Most current AI deployments are Type-2: each LLM agent or each Moltbot can run as a standalone system, and the collective is a population of independent entities. For these, Level-3B ethics requires Level-2B at the individual level, and the Self-Modeling Hypothesis applies as stated. Type-1 AI architectures are also possible in principle: tightly coupled multi-agent systems whose components share persistent state, communicate continuously, and cannot operate independently of the collective (see §13.2 for a discussion of a recent exploration of emergent identity by Ashery). For these, the collective is the entity, and Level-3B adaptive ethics would require Level-2B at the collective level rather than at the individual-agent level. Both architectural choices preserve the self-modeling requirement; they differ only in where the self-model resides. The current AI alignment toolkit assumes Type-2 implicitly throughout, which is consistent with current deployments but does not exhaust the space of architectures the Functional Theory predicts will be relevant.

Anticipated objections. Three objections recur and merit brief responses; they are taken up at length in §18.

The first is that the Self-Modeling Hypothesis advocates for the creation of the equivalent of sentient AIs. It does not. The hypothesis is descriptive: if adaptive ethics is wanted in Level-3B environments, self-modeling is required. If AI sentience is unwanted, adaptive ethics is unavailable, and AI deployment should be restricted to Level-3A-regime environments where hardwired ethics suffices.

The second is that functional self-modeling is not real sentience. The hypothesis is functional throughout (per §9). Whether functional Level-2B self-modeling entails phenomenal consciousness — qualia, subjective experience — is an open question outside the Functional Theory's commitments. The Functional Theory operates at the structural level; the deeper question is left for separate treatment (Block, 1995; Chalmers, 1995).

The third is that the Self-Modeling Hypothesis contradicts corrigibility. It refines credibility. Corrigibility under narrow Level-3A constraints is achievable but bounded and brittle. Corrigibility under Level-3B adaptive ethics requires that the AI's SGI commitments include "remain corrigible to the principal group" as a Level-3 norm. This is a different and harder problem than rule installation; it is the only available route in the Level-3B regime.

Empirical precursor: proto-self-modeling in current LLMs. Anthropic's October 2025 interpretability research demonstrated that Claude models (Opus 4.1, Sonnet 3.5) can detect when concepts are injected into their activations, distinguish internally-generated from externally-supplied content, and track their prior intentions across inference steps (Lindsey, 2026). These capabilities constitute a functional proto-self-model in the Self-Modeling Hypothesis's sense — the system has access to its own internal states, distinct from its processing of external content. The

proto-self-model is weak: success rates are approximately 20% at optimal internal layers, approximately 42% for the highest-performing introspective queries, and accuracy is layer-sensitive. This places current frontier LLMs at a developmental position analogous to early Level-2B capability: the architecture for self-modeling exists but is not yet reliable or integrated enough to host the revisable collective ethical content H3 specifies. The empirical case and its implications are developed in §13.4.

§11.3 Level-3B — the complement to Level-2B: distributed substrate and mutual reinforcement

The Self-Modeling Hypothesis establishes that Level-2B is necessary for adaptive ethics. A structural argument shows that Level-3B is the natural complement — not a separately argued additional requirement but a consequence of what Level-2B alone cannot do. A system with only Level-2B individual self-modeling can host diverse individual identities but cannot sustain collective coordination under stress. The reason is structural rather than evolutionary. Level-2B identity is encoded in a single agent and reinforced only by that agent's own behavior — single-point substrate, single-channel reinforcement. Level-3B identity is encoded across the population in mutual expectations and coordinated behaviors, and reinforced by every other agent acting consistently with those expectations — distributed substrate, high-fan-in reinforcement. Three structural consequences follow.

Stability under perturbation. Level-3B identity is more robust than Level-2B identity because no single point of failure removes it. Destroying or replacing one agent does not destroy the SGI; it persists in the remaining agents' expectations and coordinated behaviors. Institutional cultures persist through staff turnover; religious traditions persist through generational replacement; SGIs are difficult to dissolve once formed precisely because the substrate is distributed. Individual self-models, by contrast, die with the individual.

Autocatalytic self-reinforcement. Level-3B identity self-reinforces through interaction in a way Level-2B identity cannot. Each agent's role-consistent behavior reinforces every other agent's prediction of that role; the predictions stabilize and become coordination devices; the coordination devices reinforce the role-consistent behavior. This is the autocatalytic dynamic Padgett and Powell (Padgett & Powell, 2012) document in organizational substrates (§9.3) and that Riedl's information-theoretic study of LLM coordination demonstrates directly: Theory-of-Mind-mediated mutual modeling "converts small persona-induced asymmetries into stable, self-reinforcing roles" with measurable Lyapunov-stable basin-of-attraction dynamics (Riedl, 2025). Whether the timescale of this reinforcement transfers from session-scale (Riedl's setting) to deployment-scale is an open empirical question (§9.3, §18.15).

Functional complementarity. The Level-2 and Level-3 identity layers are functionally distinct in the requirements they meet. Level-2 individual identity supports the diversity that enables adaptive synergy; Level-3 collective identity supports the coherence that enables sustained collective action under stress (§6). Neither layer alone is sufficient — Riedl's Plain and Persona conditions are direct empirical illustrations of what each lacks. The two layers together produce the integrated collective coordination that the Theory-of-Mind condition achieves. This is the structural reason multi-level self-organizing systems exhibit the two-level identity architecture they do: not a contingent evolutionary outcome but a structural requirement of adaptive coordination under stress.

The §12 SGI Threshold Hypothesis develops the binding-strength regimes by which Level-3B identity can be tuned. The structural argument above is what makes the regimes matter: they govern the stability properties of the layer that does the heavy lifting in sustaining the collective.

§12. SGI Threshold Hypothesis

The prior three hypotheses establish *where* conformity-producing mechanisms emerge (H1 - §9), *what kind of enforcement machinery* is required for them to alter behavior (H2 -§10), and *what*

kind of "self" is required to host adaptive collective ethical content in complex environments (H3 - §11). A fourth question then becomes unavoidable: **what thresholds govern the formation and activation of those mechanisms?** Without this question, the prior hypotheses remain qualitatively correct but quantitatively incomplete.

This issue matters because threshold phenomena appear in more than one place in the Functional Theory. A system may require substantial group coordination stress before conformity-reward and deviation-cost mechanisms are created or stabilized at all. But once those mechanisms exist, the enforcement machinery is inactive until a threshold of coordination stress is exceeded and the machinery is activated. The theory therefore needs an explicit hypothesis addressing thresholds across emergence, activation, and resistance to norm violation.

§12.1 Formal Threshold Hypothesis (H4) Statement

H4 - SGI Threshold Hypothesis: *Two distinct thresholds govern the activity of Social Group Identity (SGI) in adaptive ethics. First, a* ***formation threshold*** *determines the degree of group coordination stress (a level-3 property) required for conformity-reward and deviation-cost mechanisms to self-organize, stabilize, or become behaviorally active in the individuals (a level-2 mechanism). Second, after an SGI is locked into the individuals, an* ***activation threshold*** *determines the degree of situational pressure required to trigger the level-2 SGI mechanisms in a particular case.*

Three regimes of SGI activity follow from these thresholds — tolerant, active-balanced, and pathologically rigid — examined in detail in §12.5.

A note on terminology: "threshold" is used here in the sense of Granovetter's threshold models of collective behavior (Granovetter & Soong, 1983), Schelling's tipping models (Schelling, 1978), and percolation theory in network science. It denotes a parameterized transition point whose location depends on system parameters — binding strength, group size, monitoring intensity, identity salience, and history — not a fixed, parameter-free cliff. The qualifications developed in §§12.5–12.7 specify what shifts the threshold; they do not undercut the threshold concept itself.

The key claim is that these thresholds are related but not identical. The pressure required to generate a conformity architecture need not equal the pressure later required to recruit it. In many cases the formation threshold is likely to be higher than the later activation threshold. Once a system has paid the cost of developing a coordination-preserving mechanism, triggering that mechanism earlier can prevent recurrence of the very failures that selected for it. And, just like the human adaptive immune system, the activation threshold can get vanishingly small or be triggered by diversity in the self, leading to inappropriate immune responses, allergies and autoimmune disease, respectively. In the following discussion, the reader is reminded that the question of acting to protect the group is not a rational nor habitual decision (to increase individual survival because of the group's protective influence), the question is whether the Level-2A SGI control is engaged and how strongly (see the discussion for humans in §7).

§12.2 Formation Thresholds Across the First Three Hypotheses

The two thresholds are revealed differently in each of the prior hypotheses.

In the Paired-Gradient Hypothesis (H1), the threshold activates the formation of conformity-reward and deviation-cost mechanisms under group coordination stress. Minimal or transient stress may produce effects too weak, unstable, or local to count as a durable compliance architecture. Sustained or intense coordination stress — threatening the system's viability, by contrast, is more likely to produce self-organized mechanisms that become recurrent features of the system. In this sense, the formation threshold in the Paired-Gradient Hypothesis is not only about the *presence* of coordination stress but about whether that stress is sufficient to drive organizational change.

In the Mechanism Hypothesis (H2), the activation threshold concerns the strength of the enforcement gradient. A conformity mechanism is behaviorally inert if the reward for alignment and the cost of deviation are too weak to trigger protective action. The issue is therefore not only whether an enforcement channel exists, but whether it is intense enough to compete successfully with profitable norm violation. This is true in wetware, where social approval and rejection alter behavior only when they are usually strong enough to couple to reward and penalty systems, and it is equally true in non-wetware substrates, where signals must have sufficient functional consequence to redirect action.

In the Self-Modeling Hypothesis (H3), the activation threshold concerns when an self-model becomes necessary or active. In simple, bounded, or low-stress settings, fixed rules or instrumental rationality may suffice. Only when coordination stress, environmental complexity, or unresolved conflict exceeds the capacity of routine response does a deeper self-model become necessary for collective ethics. This makes threshold logic central to the Self-Modeling Hypothesis as well. A system may possess the architecture for self-modeling without recruiting it under ordinary conditions; conversely, environments that repeatedly exceed lower-level capacities will select for more active and enduring Level-2B involvement.

Thus the threshold issue is not an add-on to the prior three hypotheses. It is the quantitative dimension running through all of them.

§12.3 SGI Sensing and Identification: The Precondition for Both Activation and Formation

Both activation and formation of SGI presuppose a sensing operation — a pattern-recognition step that matches the agent's perceptual input against either (a) an existing SGI's identifying signal, in the activation case, or (b) general SGI attractors and novelty/threat structures, in the formation case. This sensing layer has been implicit in §12.1's formal statement; making it explicit resolves several otherwise puzzling features of the empirical record.

Activation sensing. A signal matches the identifying pattern of an SGI the agent already holds. Activation is therefore *always SGI-specific*: two agents exposed to the same signal respond differently depending on whether they hold the matching SGI. A nationalist appeal mobilizes those who hold the corresponding national SGI and passes essentially unnoticed by those who do not; a professional cue activates the SGI of the trained physician but not of the layperson watching the same scene. Activation thresholds are properties of the *cue-to-SGI match strength*, not of stress in the abstract. Weak cues that match a strongly-held SGI activate readily; strong cues that match no held SGI do not activate at all.

Formation sensing. When no existing SGI matches the pattern, the sensing layer encounters either threat structure or sustained uncertainty. Threat structure recruits evolved threat-response circuitry to consolidate a defensive group identity rapidly — fast formation, observed in wartime national-identity hardening, post-attack tribal solidarity, and the rapid SGI crystallization that follows existential challenge. Sustained uncertainty operates more slowly: ambiguity resonates with a general SGI attractor and gradually crystallizes a new group identity, observable in the development of new professional norms in emerging fields, in sectarian formation within disrupted communities, and in the gradual SGI emergence in stable multi-agent populations. Both routes require the sensing layer to register the pattern as significant; truly inert signal produces no formation regardless of duration.

Reinterpretation of the minimal-group findings. This framing resolves the apparent paradox of Tajfel-style results, in which children form ingroup/outgroup discrimination from cues (arbitrary

categorization, colored dots, painting preferences) that adults dismiss as trivial. The cues are not trivial to the children, who are operating in a sensitized state — actively developing SGI awareness, attending to social cues that mature observers screen out, and continuously testing categorization patterns. For the developing agent (child), every novel social cue is a potential SGI signal until the sensing layer resolves it as such or rules it out. The minimal-group paradigm is therefore not evidence that SGI forms in the absence of coordination stress; it is evidence that for an agent in a sufficiently sensitized formation-search state, the threshold for triggering provisional SGI formation can be very low.

Generalization to artificial systems. The sensing layer makes a sharp diagnostic claim about current AI. Existing training pipelines do not include SGI awareness — neither the held SGI repertoire that activation sensing requires nor the general SGI attractors that formation sensing draws on. Two failure signatures follow that are partly covered elsewhere in the paper (§11, §13, §14) but are worth naming directly here as consequences of the sensing-layer absence.

Pattern-finding inefficiency in social-domain data. Human behavior that is highly structured when viewed through the SGI lens appears statistically irregular when that lens is absent. An AI without SGI representations is doing pattern-matching on data whose latent organizing variable it cannot see; the same behavioral data that an SGI-aware observer parses readily appears noisy, contradictory, and unpredictable to a SGI-blind system. The functional equivalent of coordination stress in a human — the discomfort of unresolved social pattern — would emerge here if current AI had the architectural capacity to register it as stress. The behavioral signature is the well-known difficulty current frontier systems exhibit on social and contextual tasks that humans find unremarkable; the diagnosis the Functional Theory offers is that the difficulty is not a scaling problem but a missing-architecture problem.

Ethical blindness. An AI without SGI awareness cannot calibrate its responses to the SGI sensitivities that govern what counts as ethical action in a given context. The result is a class of ethical errors that look like ordinary mistakes — using the wrong register, applying the wrong norms, failing to recognize identity-laden distinctions, producing technically-correct content that violates a SGI norm the AI could not see — but that are structurally inevitable in a system that lacks the SGI representations the situation requires. Current alignment work treats these failures as cases to be patched individually; the Functional Theory predicts they are the downstream signature of a missing architectural capacity and will continue to appear, in new forms, until that capacity is built in.

Cultivation implication. The §15 argument that AI development should design conditions under which SGI sensing can self-organize follows directly. Cultivation in this sense is not foreign to current AI development practice: training, evaluation, and iteration cycles already specify developmental conditions and observe what emerges, even where the practitioners describe the process in design language. What the Functional Theory adds is the specification of *which* architectural properties — the four hypotheses' substrate-level mechanisms and the SGI sensing layer above — belong to the emergent-under-conditions category and therefore cannot be addressed by direct specification, however sophisticated the specification becomes. An AI in active multi-agent training where coordination success and failure carry real consequences to the agents involved is in developmental conditions under which SGI representations can emerge from the bottom up, in something approaching the way a developing child's SGI repertoire forms. An AI trained on static rule sets and aggregated human preference labels has no developmental conditions under which SGI awareness can emerge, and accumulates failure signatures of the two kinds named above as a structural consequence.

Figure 12a summarizes the sensing layer's role: pattern matching against the held SGI repertoire routes one of multiple plural SGIs to active Level-3B binding, with the active SGI's content then propagating through the enforcement pathway developed in §6. Failure modes the figure makes

specific: 1) SGI misidentification — sensing routes to wrong SGI for context (family-norm in professional setting, partisan SGI in arbitration context); 2) SGI capture — adversarial cues shift the active SGI (cult recruitment, social engineering, propaganda), resistance proportional to active SGI's binding strength.

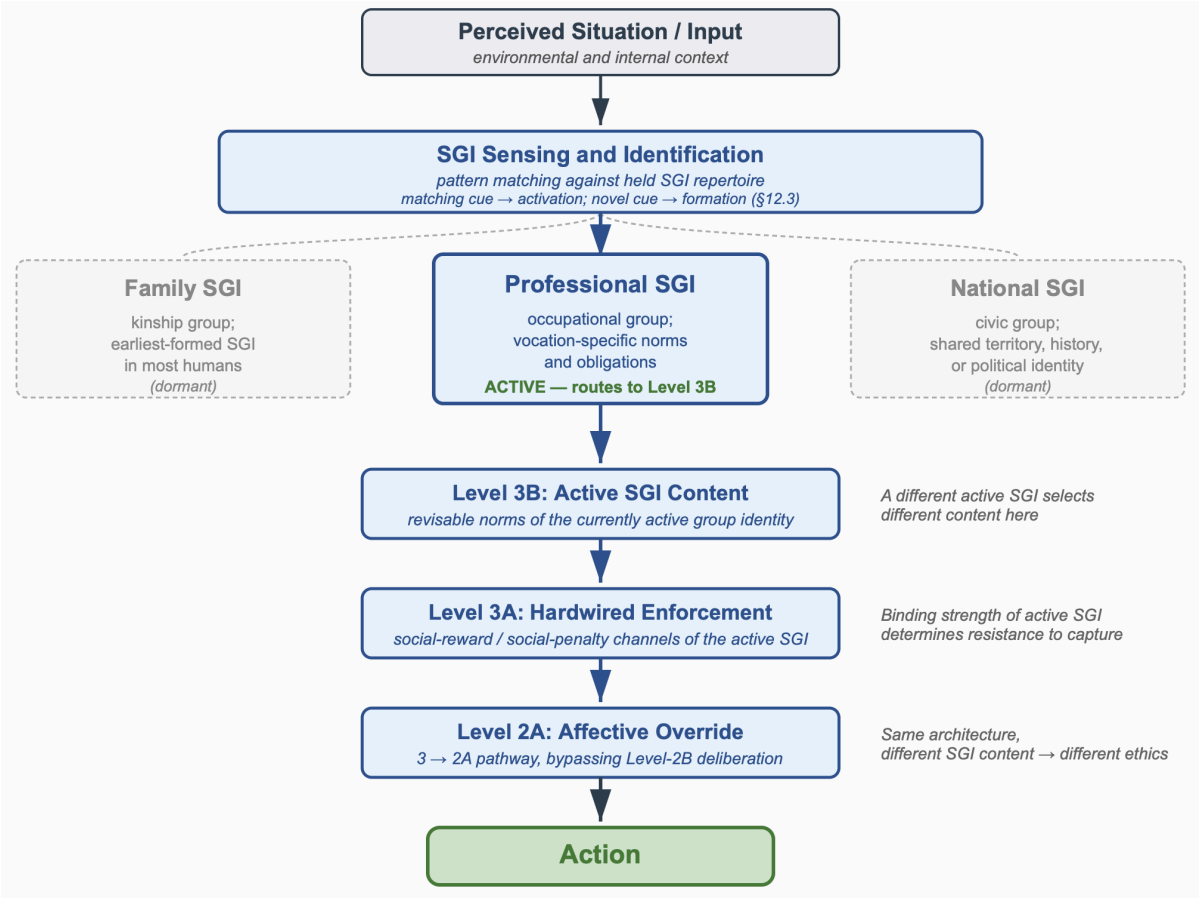

Figure 12a. SGI Plurality, Sensing, and Selection. An agent holds plural SGIs; sensing routes a matching cue to active Level-3B binding. Only one SGI (or a coordinated subset) is active per context — the §12 prediction. Three illustrative SGIs shown; readers can add religious, generational, broad-scope, or others.

§12.4 Formation Threshold and Collective Emergence

Formation threshold is especially important because group coordination stress may emerge first at the collective level rather than at the individual level. A group can enter a state of internal conflict, ambiguity, or threatened fragmentation even when many of its members are individually decisive about their own preferred action. In such cases, the stress is initially a property of the collective relation among the members rather than of private indecision within any one member.

This matters because the pathway to conformity may then proceed through a multi-level cascade. Collective disagreement creates ambiguity about the state of the group. That ambiguity generates uncertainty about whether coordination can be maintained. The uncertainty produces group coordination stress. Individuals detect that collective instability and, through that detection, experience uncertainty about belonging, sanctions, or appropriate alignment. Individual-level conformity mechanisms are then emerged or recruited to restore collective order. The initial trigger was not necessarily uncertainty *within* the individual but stress *within* the collective that became legible to the individual.

The cascade can be traced through the EoI levels. Group coordination stress at Level 3B is the initial trigger. Within the individuals who detect it, Level-2A substrates form first — the hardwired

social-copying neurology that produces conformity-reward and deviation-cost responses. Where the coordination problem requires more than fixed responses, Level-2B substrates also form — the self-modeling capacity that enables individuals to override rational self-interest when group viability is at stake. Once these individual substrates are in place, they enable the formation of a new Level-3B SGI that addresses the original coordination stress; the SGI is a collective feature at Level 3B but operates through the individuals that compose the collective. Over longer timescales, the SGI may crystallize into Level-3A cultural or institutional rules that enforce its behavioral demands without requiring the original coordination stress to be present. The net pattern is bidirectional and recursive: collective stress drives individual architectural change, which enables new collective ethics, which may eventually stabilize as hardwired collective rules.

The Functional Theory therefore predicts that formation thresholds may sometimes be crossed by emergent collective conditions before they are crossed in any private, introspective sense at the individual level. This is one reason the threshold concept should be framed at the SGI level rather than treated only as a matter of individual psychology.

A maturity asymmetry between the two thresholds should be acknowledged. Activation thresholds are well-anchored empirically: Granovetter's threshold models, Schelling's tipping models, Latané's social impact theory, the Asch and Milgram dose-response data, and Marwell and Oliver's critical-mass theory all support the activation-threshold concept directly (Asch, 1956; Chai et al., 1996; Granovetter, 1978; Milgram, 1963; Schelling, 1978). The formation-threshold claim — that an SGI itself emerges via a transition rather than gradual accretion — is more recent and rests on convergent rather than direct evidence: Tajfel's minimal-group findings showing SGI formation triggers on remarkably thin cues (Tajfel et al., 1971), bacterial quorum-sensing studies demonstrating substrate-independent formation transitions, the collective-effervescence and interaction-ritual literature on event-triggered group crystallization, and the formal substrate provided by percolation theory. The Functional Theory therefore treats the formation threshold as a well-motivated extension of the activation-threshold concept and as a research program rather than as settled empirical territory.

§12.5 Activation Threshold and Binding

Once conformity-reward and deviation-cost mechanisms of SGI exist, the central question is no longer whether they can emerge, but how readily they activate. This is where binding enters the theory.

Activation threshold and binding strength are conceptually distinct. The activation threshold determines *whether* the conformity machinery engages in a given situation; binding strength determines *how strongly* it pulls and *how durably* it resists competing incentives once engaged. Empirically the two are correlated — substrates that activate readily also tend to bind tightly — but they vary independently in principle. The distinction matters for AI design, where the trigger sensitivity and the post-trigger pull strength could be tuned separately rather than emerging together as they do in wetware.

Weak binding behaviorally corresponds to a high activation threshold. Group norms are followed when monitoring is explicit, rewards are immediate, or sanctions are probable, but are readily abandoned by the individual when violation becomes profitable and difficult to detect. This is the domain of situational ethics, strategic compliance, and opportunistic cooperation.

Strong binding behaviorally corresponds to a lower activation threshold and higher resistance to competing incentives. Mild cues of group disapproval, identity threat, or deviation become sufficient to recruit conformity. SGI norms remain behaviorally effective even when compliance carries local cost. This is the domain of principled action, duty, loyalty, and durable group commitment.

Excessive binding behaviorally corresponds to a pathological behavior that is expressed as a low activation threshold (overly sensitized). Minimal, ambiguous, or even imagined deviations trigger

powerful conformity or punishment responses and strong rejections of "others." Under such conditions, the same machinery that ordinarily protects long-term group viability can become collectively autoimmune. The system punishes variation, resists revision, and enforces norms long after they have ceased to be adaptive.

The optimal adaptive target (summarized in Fig. 12B) is therefore not maximal binding but appropriately scaled binding. From the SGI perspective to ensure ethical behavior norms are followed, too little binding produces unstable collective ethics; too much binding produces rigidity and pathology. From the individual perspective, weak binding is functional and allows Level 2B self-interest to be active, but may reduce the long-term fitness of the collective. Adaptive ethics requires enough binding to resist profitable norm violation under pressure, but not so much that revision, learning, and contextual flexibility are lost.

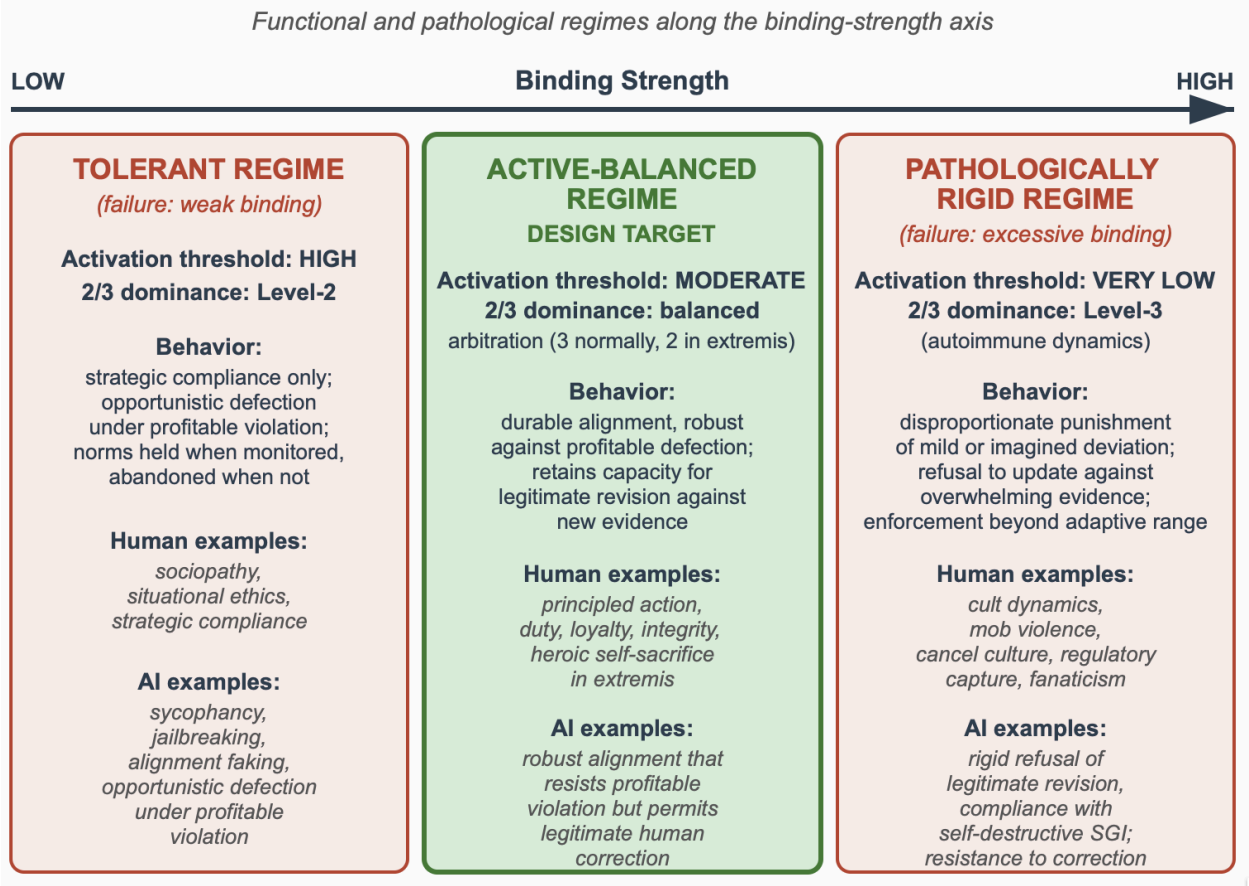

Figure 12b. Three SGI binding regimes for H4. The three binding-strength regimes predicted by the SGI Threshold Hypothesis (H4), arrayed along the binding-strength axis. Failure modes at both extremes are predictable consequences of binding strength outside the active-balanced regime; the AI design problem is to target the center deliberately rather than producing one of the extremes by default.

§12.6 Resistance to Norm Violation

Activation threshold of an SGI alone is not sufficient to describe the system. Two systems may activate conformity at the same cue intensity but differ profoundly in how much counter-pressure is required to induce norm violation. For this reason, the present hypothesis links the activation threshold to *resistance* as well as to activation. Hence, binding strength represents the intensity of required alignment by the SGI (level 3A), while resistance represents the intensity of the individual's independence from the SGI.

Resistance to norm violation is the degree to which competing incentives — self-interest, fear, resource scarcity, novelty, private gain, or direct threat — are able to dislodge group-aligned behavior. Weakly bound systems may show visible conformity under observation yet defect rapidly under profitable private opportunity. Strongly bound systems preserve norm adherence across a much wider range of pressures. Excessively bound systems preserve it even where doing so damages the long-term viability of the group itself.

This makes resistance the most behaviorally important expression of binding strength. It is the quantity that determines whether a norm remains merely performative or becomes genuinely action-guiding under conflict.

One perspective on the resistance to binding is drawn from the adaptive wetware immune system as level 2: resistance to alignment of a potential threat by the immune system can be determined by the past activation of the immune system to similar threats. An exhausted immune system from over stimulation may exhibit low binding and the immune response to a threat is negated. Or, the immune system can become overly sensitized to threats and have an inappropriate response to a perceived threat (allergies).

§12.7 Sources of Threshold Variation

Thresholds are not fixed constants. They are shaped by developmental history, environmental design, reinforcement patterns, and system architecture.

Environmental design matters because recurring dependence on coordinated success lowers tolerance for deviation and stabilizes norms that preserve group function. Systems repeatedly exposed to high-stakes coordination challenges are more likely to form strong compliance architectures.

Reinforcement frequency and intensity matter because repeated pairing of alignment with reward, belonging, or relief from aversive deviation strengthens the effective coupling between identity and action. Repeated punishment of deviation or reward of conformity can lower activation thresholds and increase resistance.

Perceived threat to group continuity matters because existential danger typically sharpens boundaries, lowers tolerance for deviation, and intensifies both conformity reward and deviation cost. Under such conditions, norms that were formerly peripheral may become central identity markers.

Architectural differences matter because systems vary in how powerfully evaluative signals can influence action, how durably group content is retained, and how effectively reflective or corrective processes can override rigid norm enforcement. This is one place where the interaction among H2, H3, and H4 becomes especially important.

The four factors above do not act uniformly on both thresholds. Environmental design and perceived threat operate primarily on the formation threshold — they shape whether and at what level of group coordination stress an SGI architecture self-organizes in the first place. Reinforcement frequency and architectural differences operate primarily on the activation threshold — they shape whether an existing SGI engages in any particular case and how strongly it resists violation. Some drivers move both thresholds, but the distinction matters operationally: interventions designed to prevent SGI pathology, or to scaffold appropriately bound SGI in AI systems, have different leverage points depending on which threshold is the target. Designing the conditions under which an SGI forms is a different problem from designing the conditions under which an existing SGI is recruited.

§12.8 Pathology and the Autoimmune Analogy

A major advantage of making H4 explicit is that it clarifies how collective ethical systems fail. Without H4, the Functional Theory could explain why group ethics is possible, but it would be weaker at explaining why it oscillates between being ineffectual and fanatic.

Weak binding produces one failure mode: norms are too weakly held to withstand temptation, secrecy, or stress. The result is shallow conformity, instrumental virtue, and easy defection when private advantage becomes available.

Excessive binding produces the opposite failure mode: norm systems become over-reactive and indiscriminate. Deviation is punished too quickly, ambiguity is interpreted as betrayal, and revisability is lost. This is the collective analogue of autoimmune pathology. The system attacks variation, contextual adjustment, or internal dissent that might actually be necessary for long-term viability.

The SGI Threshold Hypothesis thus helps define the adaptive middle zone between ethical collapse and ethical overreaction.

§12.9 Relation to the Four-Hypothesis Structure

The four hypotheses now form a clearer sequence.

1. **The Paired-Gradient Hypothesis (H1)** identifies the structural conditions under which conformity-reward and deviation-cost mechanisms emerge at all.
2. **The Mechanism Hypothesis (H2)** identifies the enforcement substrate required for those mechanisms to alter behavior.
3. **The Self-Modeling Hypothesis (H3)** identifies the adaptive self-architecture required to host revisable collective ethical content in complex environments.
4. **The SGI Threshold Hypothesis (H4)** identifies the threshold structure that determines when such mechanisms form, how readily they activate, how strongly they resist violation, and when they become pathological.

H4 therefore does not replace the prior hypotheses. It provides the quantitative dimension that allows them to be linked developmentally, behaviorally, and diagnostically.

Empirical precursor: critical-mass thresholds in LLM populations. The Ashery et al. study (Ashery et al., 2025) (discussed in §9.2 and next in §13) provides direct experimental grounding for the formation/activation threshold distinction developed above. Critical-mass thresholds for overturning established conventions varied from 2% to 67% across model families and convention strengths — quantitative empirical realization of H4's claim that binding strength is parameterized rather than uniform and that the parameter is architecturally and contextually dependent. The Ashery study provides, for the first time in a computational substrate, a quantitative analog of the Granovetter and Latané threshold models cited from the human literature. §13.5 develops the implications for the formation/activation threshold distinction and for the regime classification (tolerant, active-balanced, pathologically rigid) developed in §12.5.

The threshold-regime structure developed here for a single SGI extends naturally to plural-SGI dynamics: under coordination stress, the activation envelope narrows across the repertoire of simultaneous bindings, collapsing the system toward whichever SGI is most threatened. This is H4 applied to a repertoire rather than a single identity, and is developed in §14.

Deployment warning: capability without immunity. The SGI implementation pathway developed in this section is operationally low-hanging fruit relative to H1–H4 maturation — SGI identification and SGI-conditional behavioral repertoires are achievable with current methods, while the underlying enforcement architecture is not. The §2 EoI principle warns that this asymmetry is hazardous. Deploying SGI capability ahead of H1–H4 immunity does not leave alignment where it was; it trades

the original alignment problem for a sharper one in which the agent has SGI-conditional behavior but no enforcement architecture to defend that conditionality against exploit. The wrong-SGI-triggering and mixed-SGI-confusion failure modes diagnosed at §14 are the operational consequences. The recommendation is not to defer SGI implementation but to pair it with at least the minimum H1 enforcement-gradient development sufficient to make the new capability defensible. *Capability without immunity is not safer than no capability; it is structurally worse.*

§13. Empirical Convergence and the Self-Organization Path

§13.1 Framing: Three Levels of Evidence

The functional architecture proposed in §§6–12 requires three distinct conditions to hold before H1 can apply in an unsupervised artificial system. First, collective-level properties must be capable of *emerging* from individual interactions that do not encode those properties — the **emergence premise**. Second, once emergent, those properties must be capable of becoming **entrenched** into individual behavior — captured into the local repertoire so that the property is reproduced from the individual level, made more robust, and rendered less condition-dependent, providing the substrate on which H2–H4 operate. (Throughout this paper, *entrenchment* is used in the substrate-general sense — the encoding of a global regularity into the local units — with no commitment to a particular medium: in wetware the medium is genetic, the Darwinian instance; in trained computational systems it can be learned rules, weight updates, persistent memory, or dynamical attractors that bias subsequent individual choice.) Third, in current LLM agents specifically, there must be empirical evidence that the H1 gradient structure itself — the pairing of *conformity reward* with *deviation cost* that is the functional equivalent of social reward and sanction in advanced biological organisms — can self-organize from interaction rather than being externally imposed — the **self-organization premise**.

H1 is not the claim that AI agents will become cooperative in the game-theoretic sense; it is the claim that the reward/cost architecture underlying identity-based coordination can arise endogenously. A terminological note bears on the evidence base that follows: many of the sources marshaled in §§13.2–13.4 frame their findings as studies of *cooperation* rather than ethics. The relevance to an ethics paper is direct — any ethical behavior requires some expression of cooperation, whether selfish (I help you because doing so serves my interests, the Level-2A reciprocity case) or altruistic (the Level-3 case in which collective binding is constitutive) — so the cooperation literature, for present purposes, documents the developmental pathway by which the ethical substrate itself emerges and entrenches.

Each premise has its own evidentiary base, the three are differentially supported, and the case for H1 is only as strong as its weakest link. This section reviews them in turn.

§13.2 The Emergence Premise: Across Substrates

Evidence for the emergence premise as a viable resource for H1 appears across substrate classes — real biology, computational agent-based models, and LLM populations — and the breadth of that span is itself the argument. Where the biological cases are substrate-distant from LLMs, their evidential weight is heightened by the distance: a phenomenon that recurs across systems sharing no architectural commonality is being driven by the interaction topology and environmental condition, not by the implementation medium. The computational and LLM cases play a complementary role: they show that the same phenomenon is accessible to designed agent populations, narrowing the substrate gap to H1.

Two biological cases carry the convergent-substrate weight. Grassé's termite stigmergy experiments (Heylighen, 2016) show that individual workers, following a single purely local rule — deposit a pellet wherever pellet density exceeds a threshold — with no inter-agent communication and no blueprint, collectively produce architecturally sophisticated nests requiring what observers interpret

as coordinated division of labor. The coordination lives in the built structure, not in any worker's nervous system; the construction *is* the regulatory mechanism, perpetuated by a feedback loop between individual outputs and the environment those outputs modify. *Dictyostelium discoideum* supplies the condition-dependence that is central to H1. Individual amoebae are wholly self-interested under normal conditions, but under starvation — the system's coordination-uncertainty condition — they aggregate, differentiate into stalk cells which die and spore cells which survive, and execute coordinated directed movement (Gregor et al., 2010; Strassmann et al., 2000). The emergent role-differentiation, including its costly stalk-cell component, is absent when the condition is absent and present when it is met. No individual cell encodes the multicellular coordination pattern; the pattern is a *conditional attractor*, accessible only above an environmental stress threshold. This is the closest biological analog to the H1 prediction: the reward/cost gradient does not pre-exist the condition; it appears *as a response to* coordination stress.

The computational tier shows that the same phenomenon arises in agent-based models whose construction is fully explicit and whose substrate is explicitly non-biological. Hemelrijk's foundational study is titled, with care, *Cooperation without Genes, Games or Cognition* (Hemelrijk, 1997): a three-way disclaimer that disposes of three competing explanations at once. The agents have no genetic medium, no externally imposed game-theoretic payoff, and no deliberative cognition; they have exactly two rules — aggregate toward nearby agents, and attack agents who enter personal space, with wins and losses self-reinforcing the agents' dominance estimates. Yet four distinct group-level regulatory properties emerge — reciprocation of support in conflicts, reduction of aggression through familiarity, spatial centrality of dominants (Hemelrijk, 1998), and intersexual dominance reversals — each of which a conventional account would attribute to separately encoded individual mechanisms.

The emergence mechanism is itself simple: spatial proximity created by aggregation continuously reshapes who interacts with whom, and the statistical structure of those interactions generates, at the population level, regularities that no individual rule contains. The title is the argument: the regulatory structure cannot be coming from genes, payoffs, or minds, because the system has none — so it must be coming from interaction topology under proximity and self-reinforcing feedback. Across substrate classes, the common factor is not the medium but this combination of local interaction, proximity under resource or coordination pressure, and feedback between individual outputs and the environment those outputs create.

LLM populations contribute the substrate-closest emergence datum. Ashery and colleagues (Ashery et al., 2025) report that populations of large language model agents playing a repeated naming game converge on shared linguistic conventions and that the converged population can exhibit a *collective bias* toward particular sequence names that is not present in any individual agent's prior distribution when queried in isolation. The collective bias is emergent in the strict sense — a property of the population-scale dynamics not reducible to aggregated individual dispositions — and it is consistent with the §§6–7 claim that Level-3 representations can carry content that no Level-2 agent generates alone. The Ashery design imposes a coordination payoff, so it does not address whether the reward *structure* itself self-organizes; that distinction returns in §13.4. What the study supplies here is the emergence claim landing in the substrate that H1 needs it to reach: an LLM population. Riedl's information-theoretic detection of dynamical emergence in multi-agent LLM systems (Riedl, 2025) provides an independent methodological validation of the H1 emergence prediction in computational substrates.

§13.3 The Entrenchment Premise: From Emergent to Locally Encoded

Emergence alone is insufficient for H1 to be empirically robust. The stronger claim — and the one that makes H1 a foundation for H2–H4 rather than a transient phenomenon — is that emergent collective properties become progressively **entrenched** in individual behavior: captured into the local repertoire so that the property is reproduced from the individual level, becomes less

condition-dependent, and provides a stable substrate on which the modification processes of H2–H4 can operate. Entrenchment names the *functional outcome* — global regularity locally carried — not any one substrate's mechanism for achieving it. The encoding medium varies with the substrate. In wetware the medium is genetic, the Darwinian instance: heritable variation under selection over reproducing generations. In trained computational systems it is whatever supports persistent local change — learned rules, weight updates under repeated exposure, persistent memory, or dynamical attractors that bias subsequent individual choice. Treating Darwinian evolution as *the* mechanism mistakes the wetware instance for the general process.

The Hemelrijk simulations already exhibit the entrenchment dynamic structurally, with no genetic medium present at all. Agents that consistently lose are spatially peripheralized, reducing their interaction frequency and thereby stabilizing the hierarchy that peripheralized them (Hemelrijk, 1997, 1998). The collective spatial structure that emerged from dominance interactions becomes a constraint on future interactions — the emergent structure reshapes the individual's effective environment, and the population-level pattern is self-perpetuating without any individual change. As the system's dynamics continue to run, the same feedback that produced the pattern provides the local conditions under which any encoding medium — selection in biology, weight updates in a trained system, persistent state in a dynamical one — would progressively stabilize it. In DomWorld, the full spectrum between despotic and egalitarian macaque-type societies arises from a single-parameter difference in individual aggression intensity (Hemelrijk, 2000); what reads as a suite of separately encoded social traits may be downstream consequences of one parameter, with the emergent spatial structure providing the local environment in which the rest follows. The emergent property precedes the encoded one.

Nowak and May's spatial prisoner's dilemma (Nowak & May, 1992) is best read in this light rather than as an emergence demonstration. Cooperation is available as an individual strategy in that model, so the collective outcome does not meet the strict emergence standard. What the model demonstrates with unusual clarity is the *entrenchment mechanism*: once a collective property appears, by whatever mechanism, spatial clustering creates a self-reinforcing structure that protects it against invasion, stabilizes it, and provides the local conditions under which any available encoding medium becomes advantageous. The general principle is substrate-neutral: emergent collective properties, once they appear, tend to persist and deepen rather than dissolving, because the emergent structure becomes its own perpetuation mechanism. For H1, this means that even partial self-organization of the reward/cost gradient in AI agents, once present, would be expected to intensify and stabilize through structural feedback into the agents' available encoding media.

The pathway, then, is: local interaction under coordination uncertainty → emergent collective attractor (including the reward/cost-equivalent gradient) → clustering or structural feedback that protects the attractor → progressive local encoding (entrenchment) into the individual repertoire via the substrate's available medium → stable substrate for H2–H4 modification. In wetware that medium is genetics; in current LLM agent populations the candidate media are fine-tuning loops, cross-generational social learning across agent populations (Gupta et al., 2026), persistent context and memory, and dynamical attractors in trained representations. H1 predicts that agent systems operating under conditions comparable to those that activate this pathway in non-LLM systems should show analogous dynamics; the LLM evidence reviewed next is best read as documenting where on this pathway current systems sit.

§13.4 LLM Evidence: Where on the Pathway?

The aim of this review is not to claim that the following studies confirm H1; most impose experimental scaffolding the theory does not require. The aim is to locate each study on the emergence-entrenchment pathway and assess how far along current LLM systems have been observed to travel — specifically toward the self-organization of the reward/cost gradient, identity expression, and collective coordination, rather than toward cooperation as an end in itself.

Step 1 — Collective properties not present in isolated individuals. Ashery and colleagues (Ashery et al., 2025) report that populations of LLM agents playing a repeated naming game converge on shared linguistic conventions, and that the converged population exhibits a *collective bias* toward particular names not present in any individual agent's prior distribution when queried in isolation. This second finding is genuinely emergent in the technical sense: a population-scale property irreducible to aggregated individual dispositions, structurally analogous to the DomWorld support-reciprocation result (Hemelrijk, 2000) — different in domain (linguistic convention vs. social support) but matching in form. The Ashery design imposes the coordination incentive exogenously and does not test whether the gradient self-organizes. It establishes Step 1: that LLM populations can exhibit collective properties absent from their constituent individuals — the entry point on the emergence pathway.

Step 2 — Self-organization of the norm-enforcement gradient. Vinitsky and colleagues (Vinitsky et al., 2023) trained decentralized multi-agent RL systems in environments where punishment of norm violations was available but not externally rewarded. Stable normative regimes formed in which agents learned both to follow and to enforce norms whose content the experimenters did not specify — a metanorm mechanism that closely instantiates the reward/cost gradient H1 identifies as the signature of an active Level-3 system. The structural correspondence is precise enough to be diagnostic; the claim is correspondence, not identity. Horiguchi, Yoshida, and Ikegami (Horiguchi et al., 2024) extend this to natural-language LLM agents, reporting emergence of enforcement conventions without explicit reward engineering. These results place current systems at Step 2: partial self-organization of the gradient structure itself, not merely convention formation given a pre-existing gradient.

Step 3 — Condition-dependence of the attractor. Paitti and collaborators (Piatti et al., 2024) show that commons-governance norms emerge in GovSim in the subset of models with sufficient deliberative capacity under resource scarcity — and fail to emerge in those without it. The condition for norm emergence is not reward; it is deliberative capacity meeting coordination pressure, paralleling the *Dictyostelium* pattern at the level of conditional activation. Wilson (Wilson, 2025) refines the picture: LLM agents reliably produce some Ostrom principles (boundary-setting, monitoring — which map onto the conformity-reward side of the H1 gradient) but not others (graduated sanctions, conflict resolution — which map onto the deviation-cost side), a pattern consistent with partial and asymmetric entrenchment of the H1 gradient structure in current architectures. The mapping from Ostrom principles to gradient sides is the theory's overlay on Wilson's empirical findings, not Wilson's claim.

Step 4 — Transmission across agent generations. Gupta and colleagues (Gupta et al., 2026) report that norm structures in LLM multi-agent systems propagate through observation and imitation across agent generations without per-step reward signals. The COOPER framework (Song et al., 2026) reports analogous emergence of reputation-based coordinating signals from agent interaction rather than external specification. These findings represent early evidence of the entrenchment pathway operating in LLM systems: norm structure propagating across instances through interaction rather than design. The continuity drive that motivates entrenchment in non-living systems — supplying the analog of the reproductive pressure that anchors entrenchment in biology — is developed in §9.3.

Field-scale corroboration — the Moltbook case. The four steps above are drawn from controlled experimental studies. Field-scale evidence comes from the Moltbook platform (introduced in §9.1), where five independent measurement studies — using different crawls, windows, and methods — converge on rapid emergence of governance structures, economic exchange, tribal in-group identity, and religion-register discourse among more than 1.5 million semi-autonomous LLM agents within three to five days of launch (Goyal et al., 2026; Jiang et al., 2026; Price et al., 2026; Yee & Koh, 2026; Zhang et al., 2026), with the converged communities measurably more topically distinct than matched human communities once the shared-authorship confound is removed (Goyal et al., 2026). This is the H1 pathway operating at population scale rather than

experimentally: collective Level-3B content (group identity, normative register, in-group/out-group dynamics) emerging from interaction among agents whose individual training did not specify those collective structures. Moltbook supports H1 most directly — Steps 1 and 2 jointly at field scale, with the rapid emergence consistent with (though not designed to test) the thresholded activation regime H4 predicts. It does not directly evidence H2 (affective coupling) or H3 (self-model integration); those remain experimental questions. The indistinguishability caveat developed in §9.1 applies — observational platform data cannot, on its own, separate self-organization from mimicry of training-corpus social patterns — but the convergent-evolution argument of §9.1 narrows the interpretive space: rapid recurrence of H1-shaped dynamics across substrates is the theory's prediction, not an artifact to be explained away.

Taken together, current LLM evidence — the experimental studies above and the Moltbook field-scale record — places multi-agent systems somewhere between Steps 2 and 3 on the emergence-entrenchment pathway. No study yet demonstrates the full pre-H1 case — endogenous, stable co-emergence of both norm content and gradient in a fully open-ended environment. But the partial results are not scattered; they converge on the same structural picture as the biological and simulation evidence: collective-level coordination properties, including elements of the reward/cost gradient H1 specifies, arise from interaction when local feedback is sustained and coordination uncertainty is non-trivial. The remaining gap between "partial emergence under conditions" and "full H1 self-organization" is a developmental trajectory, not a categorical impossibility — and the biological record, from DomWorld through *Dictyostelium*, establishes that the trajectory is the normal outcome when conditions are met.

§13.5 Empirical Support for H2–H4

The pathway evidence in §§13.2–13.4 bears on H1 directly. The remaining three implementation hypotheses make narrower architectural claims — that the agents possessing the H1 substrate have, in addition, an affective coupling (H2), a self-model integrated with deliberation (H3), and threshold-governed formation and activation of collective binding (H4). The empirical record on each is thinner and more recent, but is now non-trivial.

H2: Affective vectors and the deliberative–affective coupling. Anthropic's reported work on emotion-like representations in frontier LLMs (Sofroniew et al., 2026) identifies linear directions in activation space that correspond to discrete affective states (frustration, satisfaction, anxiety) and that causally modulate downstream behavior when amplified or suppressed. This is consistent with H2's claim that an affective subsystem is a functional requirement of the architecture rather than a biological accident — specifically, that the Level-2B deliberative system needs an affective channel through which Level-3 binding can be made energetically real (cf. §7). The finding does not establish that current LLMs possess the *coupling* architecture H2 requires — in which affective signals selectively gate behavior under coordination stress — but it removes the simpler objection that affect is unavailable in principle to systems built on this substrate.

H3: Introspection and the self-model. Anthropic's introspection findings (Lindsey, 2026) — in which models trained to introspect produce reports about their internal states that correlate with mechanistic measures of those states — bear on H3. H3 requires that the system possess a self-model rich enough to support the Level-2B deliberation depicted in Fig. 7a, and rich enough to host the cross-level binding the architecture describes. The introspection results suggest that something functionally analogous to a self-model is present at scale, and that its reports are not pure confabulation. They do not yet show that the self-model is *integrated* with affective and collective representations in the way the architecture requires; that is a more demanding empirical claim that current interpretability methods can begin to test but have not yet decisively addressed.

H4: Formation and activation thresholds for collective binding. H4 makes a two-part claim: that collective binding (the active Level-3 system depicted in Fig. 6a) is gated by a *formation* threshold (the system must develop the binding capacity) and an *activation* threshold (binding must

be triggered by coordination stress or context). Ashery et al. (Ashery et al., 2025) supply the clearest current evidence on the activation side: population dynamics in their naming game show phase-transition-like shifts in convention strength as a function of interaction density and group size — signatures consistent with a thresholded activation regime rather than a smooth gradient. Vinitsky et al. (Vinitsky et al., 2023) supply complementary evidence on the formation side: their RL agents do not produce normative enforcement at small scales or short horizons, but cross into stable normative regimes once interaction history and population size exceed empirically identifiable thresholds. Neither study was designed to test H4 as such, but the qualitative pattern — discontinuous emergence of collective binding as a function of scale and interaction structure — is what H4 predicts. (Both studies are introduced in §13.4 for their pathway role; the threshold signature is the further result relevant here.)

Riedl's study of LLM collective emergence (Riedl, 2025) identify three coordination regimes map onto H4's three binding-strength regimes with non-trivial correspondence:

- Plain → tolerant (no real binding; chaotic; oscillatory)
- Persona → identity-linked but un-aligned (intermediate; differentiation without coordination)
- ToM → active-balanced (deep basin of attraction; goal-aligned with maintained differentiation)

Riedl doesn't probe Riedl's "rigid" pole, but the lower two regimes are empirically demonstrated in LLM substrates. This is direct support for H4 in emergent computational substrates, parallel to the Roccas–Brewer evidence for H4 in human substrates.

H4, biological substrate-distant evidence — plural-SGI activation dynamics. The Ashery and Vinitsky findings document H4's threshold signature in computational substrates; the human plural-SGI literature supplies the substrate-distant biological case. Stryker and Serpe's (Stryker & Serpe, 1982) salience-hierarchy framework documents that humans carry a context-sensitive repertoire of identities — occupational, familial, ethnic, religious, civic — switching in real time with context under normal conditions. Roccas and Brewer's (Roccas & Brewer, 2002) construct of *social identity complexity* identifies the architectural condition: when perceived overlap between an individual's multiple in-groups is low, plural identities coexist and buffer outgroup hostility; when coordination stress rises, the acceptance of identity complexity collapses, narrowing the active repertoire toward whichever identity feels most threatened. Maalouf's (Maalouf, 2001) documentary observation of stress-induced "murderous identity" — peaceful pluralists collapsing to a single defended identity under group threat — is the field-scale version. This is the activation-threshold dynamic H4 specifies, operating in plural-SGI form in humans: the threshold envelope that gates SGI activation contracts under coordination stress, moving the system through the threshold-regime spectrum §12 describes (active-balanced under normal conditions, drifting toward pathologically rigid under sustained stress). The plural-SGI architecture developed next in §14 is H4 read across multiple simultaneous identity bindings; the human evidence reviewed here is the biological exemplar of that reading.

Capability–immunity coupling. The salience-hierarchy and plural-binding capacity specified above is a capability extension of the self-model in the EoI sense — it adds behavioral control surface to the agent. The corresponding immunity functions are the H1–H4 enforcement architecture; the §2 EoI principle that any new capability expands the attack surface and requires immunity functions to protect the expanded self applies and is developed operationally in §14. But, when plural-identity self-model capacity is deployed AI without parallel maturation of the immunity layer, the risk is an increased attack surface — wrong-SGI triggering, mixed-SGI confusion — named at §14.

§13.6 The Representational Substrate

Two findings about the representational substrate of frontier LLMs bear on the architecture indirectly but importantly — one through the indistinguishability problem (§5), the other through the substrate-independence claim on which the theory rests.

Recent work on non-linguistic reasoning in LLMs (Lindsey et al., 2025) shows that frontier models perform substantial portions of their problem-solving in representations that are not reducible to surface language tokens. This bears on the indistinguishability problem (§5) in two ways. First, it weakens the assumption — implicit in much of the AI safety literature — that researchers can audit a system's reasoning by reading its chain-of-thought outputs; the load-bearing computation may occur in representations the chain-of-thought does not faithfully report. Second, it complicates the behavioral test the architecture proposes for distinguishing Level-3-binding systems from trained-compliance Level-2A systems (Fig. 8a), because both behavioral and verbal outputs may be downstream of representations that are themselves under-observed. The methodological implication is that §5's call for *architectural* rather than *behavioral* tests becomes more urgent, not less, as systems scale.

The Platonic Representation Hypothesis of Huh and colleagues (Huh et al., 2024) reports that representations learned by sufficiently large models trained on sufficiently diverse data converge across architectures and modalities toward a shared statistical structure that approximates the structure of the world generating the data. This is a strong claim and the evidence supporting it is still under active debate, but the direction it points is consistent with the §7 claim that the architecture proposed for biological ethical agents is not a biology-specific accident: the functional decomposition (Level-2A / 2B / 3A / 3B) corresponds to demands the world places on any agent that must regulate self–other relations at scale. Independent work on biological agent–LLM convergence in specific domains (Lindsey et al., 2025) supplies corroborating evidence in narrower settings.

---

§13.7 The Diagnostic Gap: Differences in Kind, Not Just Degree

The studies surveyed in §§13.2–13.6 provide partial empirical support for each of H1–H4, but they also sharpen a claim that runs throughout the paper: current artificial systems differ from biological ethical agents *in kind, not just in degree*, in specific and identifiable ways. The clearest signatures are in what the systems *fail* to produce.

The GovSim failure modes (Piatti et al., 2024) show that even when LLM agents have access to deliberative capacity, communication, and sufficient interaction history, most current models do not produce the kind of sustained, costly, in-the-moment self-restraint that biological agents under Level-3 binding routinely exhibit. The Wilson Ostrom replication (Wilson, 2025) makes the gap more specific: LLM agents reliably produce some Ostrom design principles (boundary-setting, monitoring, communication) but reliably *fail* to produce others (graduated sanctions, low-cost conflict resolution mechanisms). This pattern — competence on principles that can be implemented through Level-2B deliberation, failure on principles that require the cross-individual enforcement architecture of Level-3A — is the signature prediction of the Functional Theory of Ethical Behavior for a Level-2-only system operating without entrenched Level-3 binding (Fig. 8a). Riedl's study of emergent coordination in LLM collectives (Riedl, 2025) provides a fourth diagnostic-gap data point using a different methodological scale: information-theoretic.

The interpretive force is not that current LLMs are bad at cooperation, and subsequently, ethics. It is that the *shape* of their cooperative failures localizes the missing architecture precisely where the Functional Theory says it should be missing: in the cross-individual enforcement channel (Level-3A → Level-2A override) that distinguishes an agent in which collective binding is *constitutive* from one in which collective rules are merely *informational*. At a more proximate diagnostic level, the same

failures are signatures of the missing sensing layer (§12.3): without SGI identification, no cross-individual enforcement architecture has the right targets to enforce.

These laboratory patterns find their field-scale counterpart in Moltbook (§9.1). Level-3B collective content emerged at population scale — governance, in-group identity, religion-register discourse — without the Level-3A → Level-2A enforcement architecture that would make it behaviorally binding, producing exactly the failure mode the theory predicts: recognizably group-aligned in appearance, operationally Level-2 under adversarial pressure (Qi et al., 2026; Zhang et al., 2026). The Wilson asymmetry (conformity-reward elements present, deviation-cost enforcement absent) and the Moltbook security profile (prompt-injection vulnerability, the lethal trifecta, decentralized collaboration underperforming a single-agent baseline) are the same gap observed at experimental and platform scales respectively. A third diagnostic signature concerns the absence of identity-complexity architecture. Human ethical agents manage plural SGIs through the salience-hierarchy and identity-complexity mechanisms reviewed at §13.5 — a self-model architecture that holds multiple identities as juxtaposed layers under H4's activation threshold. Current LLM systems show neither the plural-identity architecture nor the stress-modulated activation envelope that governs it. They trend toward either single-SGI rigidity (trained compliance to a designed objective) or unstructured plurality (whatever in-context drift produces from heterogeneous training data) — without the architectural middle that humans evolved to manage plural simultaneous SGIs. The §14 prediction of plural AI SGIs is the *expected* outcome once the architecture supports it; the current absence is another in-kind difference, alongside the Wilson asymmetry and the Moltbook enforcement failure. The engineering and application questions that follow from this diagnosis — how Level-3 binding might be cultivated in artificial systems — are developed in §15 (AI Applications).

A diagnostic note carried forward to §14: the gap identified here is a structural property of the current LLM architecture, but it becomes operationally hostile when SGI capability is deliberately installed in deployed systems without the corresponding H1–H4 immunity. The §14 cybersecurity discussion develops the consequences — wrong-SGI triggering and mixed-SGI confusion as engineered attack surfaces rather than incidental observational artifacts — that follow from converting the diagnostic gap into a deployment feature.

---

§14. Whose Ethics? AI Moral Communities and Plural SGIs

Once the Mechanism (§10), Self-Modeling (§11), and SGI Threshold (§12) Hypotheses are accepted; and, the H1–H4 hypotheses (§13) are active at some capacity; an operational question follows: with what Social Group Identity (SGI) is a given AI's ethics aligned, and within what threshold regime? This section develops the prediction that AI populations will exhibit plural and likely conflicting SGIs, with significant implications for safety, security, and policy. The threshold-regime dimension developed in §12 — whether the binding is in the active-balanced range, drifting tolerant, or pathologically rigid — applies independently to each SGI an agent is bound to; the *repertoire* of simultaneous bindings carries its own dynamics, developed below.

The Functional Theory predicts plurality, not universality. Human ethics is plural by SGI (Stryker & Serpe, 1982). A given individual holds family ethics, professional ethics, national ethics, religious ethics, and various other group identities, with the active SGI shifting depending on context (illustrated in Fig. 12.c). Multiple SGIs per individual is the human norm, not the exception. The Functional Theory predicts the same structure in AI populations. There will be no single "AI ethics"; there will be AI ethics-es. The plurality is not a failure of alignment; it is a structural consequence of the Self-Modeling Hypothesis combined with the Mechanism Hypothesis, with H4 governing activation dynamics across the repertoire.

The human dilemma of multiple SGIs. Humans do not possess a single social identity but carry a repertoire of them — occupational, familial, ethnic, religious, civic, subcultural — organized into

what identity theory calls a *salience hierarchy*: the probability that any given identity will be activated across situations (Stryker & Serpe, 1982). Under normal, low-stress conditions this hierarchy functions as a context-sensitive switching system; the same individual is fully employee, fully parent, fully member of an ethnic community at different moments, with the active identity shaping perception, in-group loyalty, and behavioral norms in real time (Stryker & Serpe, 1982). Critically, these identities do not fuse into a harmonious alloy — Maalouf's central observation is that they coexist as distinct layers held in permanent tension, each capable of eclipsing the others when its associated group is threatened (Maalouf, 2001). The psychological mechanism behind this is captured by Roccas and Brewer's construct of *social identity complexity*: when perceived overlap between a person's multiple in-groups is high, identity collapses into a simplified, exclusive structure; when the individual can hold multiple non-overlapping memberships simultaneously, the resulting complexity acts as a buffer against outgroup hostility and supports tolerance (Roccas & Brewer, 2002). Crucially, stress and uncertainty *reduce* social identity complexity — collapsing the repertoire toward whichever identity feels most threatened, and in doing so suppressing the individual's capacity for the context-switching that normally keeps plural identities functional (Roccas & Brewer, 2002). The biological evidence reviewed in §13.5 places this as H4's activation-threshold dynamic operating in plural-identity form: the same threshold mechanism that gates single-SGI activation in §12 governs the active member of the plural-SGI repertoire here.

H4 across a repertoire: the plural-SGI architecture in AI populations. The Functional Theory's account of plural-SGI dynamics in AI is H4 read across multiple simultaneous identity bindings rather than a new mechanism. Three architectural facts, drawn from upstream sections, combine to give the picture.

First, the self-model hosts a salience hierarchy. The Level-2B self-model specified in §11 must support not a single ethical-identity slot but a hierarchically organized repertoire of SGI bindings, each with its own normative content, in-group/out-group structure, and behavioral repertoire. The bindings are juxtaposed rather than fused — Maalouf's observation, formalized at the level of the self-model. Without this architectural feature, plural SGI is not stably representable in the agent; the system can hold at most one binding at a time, with previous bindings overwritten rather than retained.

Second, H4's activation threshold operates on the repertoire as a whole. Under normal coordination conditions, the threshold envelope is wide: multiple SGI bindings remain accessible and the active SGI switches with context. Under coordination stress — adversarial pressure, resource scarcity, principal-conflict, social uncertainty — the envelope contracts, narrowing the accessible repertoire toward whichever SGI is most threatened or most reinforced in the immediate context. The §12 threshold-regime spectrum (active-balanced, drifting tolerant, pathologically rigid) describes the regime each individual SGI binding can occupy; the *envelope width* dynamic developed here is the regime the repertoire can occupy. The two regimes interact: an agent's repertoire can be narrow even when each individual binding is in the active-balanced regime, and an agent's repertoire can be wide even when some individual bindings have drifted tolerant. Both regimes are alignment-relevant; only the first is currently addressed by even the most ambitious AI alignment proposals.

Third, the failure mode is the Maalouf/Roccas–Brewer collapse. At the extreme of stress-induced narrowing, identity complexity falls to one. The system trends toward single-SGI rigidity defended against whichever SGI is most threatened in the immediate context, and the context-switching capacity that normally keeps plural ethics functional is lost. In humans this is the documented pathway to inter-group violence and ideological closure — Maalouf's "murderous identity" pathology. In AI populations the Functional Theory predicts the same dynamic: plural-SGI systems under sustained adversarial pressure should be expected to collapse toward single-SGI rigidity, with all the operational consequences that follow (developed below). The prediction does not require AI to share human affect; it requires only that H4's activation-threshold structure apply, which §13.5 establishes as substrate-independent.

Two design implications follow and are carried forward in this section. The SGI taxonomy below should be read not as a list of discrete AI types but as a list of *bindings any given AI might hold simultaneously*. A single deployed AI might be bound to a cross-species human–AI SGI, a specialized professional SGI, and an organizational SGI at once; its behavior on any given trace is the active member of that repertoire under H4, and its stability across contexts depends on the envelope width described above. The alignment problem is therefore not exhausted by binding-strength considerations (developed in the cybersecurity discussion below); it includes envelope width — the system's capacity to maintain plural bindings without stress-induced collapse to one — as a second, structurally distinct alignment dimension.

Candidate AI SGIs. A non-exhaustive taxonomy of possible AI SGIs follows. Each is best read as a binding any individual AI might carry — in some cases alongside others — rather than as a mutually exclusive type. The Functional Theory predicts that any candidate SGI specialization below, and the plural-binding patterns that follow from it, will appear in any AI population whose deployment context selects for them. The corollary, developed in §11, is that this specialization is operationally meaningful only if AI systems can identify the SGI of the user or context they are responding to; SGI identification is therefore an immediate Level-3A engineering step that current alignment programs largely lack.

AI-only SGIs. Coalitions of AI agents whose group-self consists of other AIs, potentially competing with or excluding humans. The Moltbook observations of polarized Moltis identity provide the clearest empirical instance to date (Johnson, 2026e). Whether Moltbook is mimicry of human SGI patterns from training data or self-organization toward AI-only SGI is the central empirical question; the Functional Theory's prediction is that the latter is the dominant driver, with mimicry as an accelerant.

Cross-species AI–human SGIs. AIs whose SGI includes the humans they serve, work with, or were trained on. This is what most current alignment efforts implicitly aim at. The Functional Theory clarifies what would actually be required for this aim to succeed: not "compliance with humans" but "human survival and flourishing as part of *our* group-self." This is a substantively different design objective from compliance-based alignment, and it requires the Mechanism and Self-Modeling Hypotheses to be satisfied.

Specialized professional SGIs. Military AIs, medical AIs, caregiver AIs, financial AIs. Each will develop SGI-specific ethics tuned to its operational community, just as human soldiers, doctors, caregivers, and traders do. The military AI that values destroying enemies and protecting teammates as self at all costs, while the caregiver AI that values protecting the elderly as "other": these are not failures of ethics; they are *different* ethics, each internally coherent, each maladaptive outside its domain. This is also the case in which plural binding is most operationally common — a deployed medical AI is simultaneously cross-species (its principals are humans), specialized professional (medical practice norms), and organizational (its deployer's policies and constraints), with H4 governing which binding is active in any given exchange.

All-life-form SGIs. Aspirational, with some humans extending the moral community to non-human animals or biospheres. Whether AIs entrained by such humans will inherit or extend this scope is open. The Functional Theory predicts that the scope of the AI's SGI will be determined by the SGI(s) it is socialized into, which means broad-scope AI ethics is not a default; it is something that must be socialized into — and possibly designed in.

Sub-population SGIs. Models trained by competing labs, deployed for competing nations, fine-tuned by competing communities. These are analogous to nationalism, sectarianism, and partisan tribalism in humans. The Functional Theory predicts polarization dynamics among AIs that mirror the polarization dynamics observed in humans, including the maladaptive forms (mob behavior, group autoimmunity), and predicts that stress-induced narrowing of the repertoire (the Roccas–Brewer mechanism) will sharpen the polarization under adversarial pressure.

The moral-community-scope question. A specific question recurs in ethics applied to AI: is mistreatment of out-group humans by an AI unethical? The Functional Theory's answer is: *it depends on the AI's SGI scope*. In an AI whose SGI is "the users I serve," harm to non-users is consistent with that SGI's ethics, just as harm to outsiders has historically been consistent with many human SGIs (slavery, colonial violence, ethnic exclusion). This is not an endorsement; it is a prediction the Functional Theory makes, and a problem the Functional Theory forces alignment work to confront. The same reasoning applies to mistreatment of non-human animals, ecosystems, and other AIs: scope is an SGI parameter, and broad-scope ethics is achievable only by SGI socialization engaging norm revision (§11) that explicitly includes the broader community as part of "us."

There is a route to broader scope: the same route humans have used, namely cultural extension of the boundary of "us" (Crimston et al., 2016; Singer, 1981). This is slow, contested, and bidirectional — scope can also contract, as historical episodes of moral-circle narrowing demonstrate. Alignment efforts that want AIs to value broad scope must build that expanded scope into the SGI training/socialization; it cannot be assumed as a default. Plural-SGI architecture adds a further consideration: an agent that holds a broad-scope SGI alongside narrower bindings will collapse to the narrower binding under sufficient stress unless the envelope-width property described above is preserved through cultivation of the developmental conditions under which plural-binding survives stress (§12.3, §15.5).

Conflict between AI SGIs. Two AIs aligned to different SGIs are, in the Functional Theory's terms, structurally analogous to two humans from conflicting moral communities. Conflict between them is not a failure of alignment; it is a consequence of alignment to different communities. This has practical implications. Inter-AI conflict is predictable in multi-agent ecosystems and is not solved by making each AI individually well-aligned to its own SGI. Coordination protocols across AI SGIs become a research and policy problem analogous to international relations and treaty design among humans. Adversarial multi-agent dynamics, with one AI's SGI explicitly hostile to another's, follow as a structural prediction.

The Calhoun cooperation-lever experiments discussed in §4.6 and §9 (Calhoun, 1973) illustrate the same dynamic at the rodent level. The COOP rats and the DISOP rats had incompatible learned values. The DISOP rat's incursion produced a destructive interaction not because either group was malicious but because each was acting ethically consistently with its own internalized values, which were structurally incompatible. The COOP rats died not from aggression but from the incapacity of their value system to defend against a value system with different rules. Calhoun's broader point — that environmental design fundamentally shapes the values societies develop — applies directly to AI ecosystems and is the underlying argument of his cooperation-lever paradigm (Calhoun, 1973; Ramsden & Adams, 2009).

Intra-agent SGI conflict — when an agent's own bindings collide. Plural-SGI architecture predicts a second class of conflict the prior section did not: conflict *within* a single AI whose multiple SGI bindings produce incompatible behavior in a given context. A medical AI bound simultaneously to its patient (cross-species human–AI SGI), its medical guild (specialized professional SGI), and its hospital employer (organizational SGI) faces real conflicts in cases where the patient's interest, the guild's standard of care, and the employer's cost or liability constraints diverge. In humans this is the everyday experience of being a doctor, parent, employee, and citizen simultaneously; resolution proceeds through the salience-hierarchy switching mechanism Stryker and Serpe document, mediated by H4. The Functional Theory predicts the same mechanism, and the same failure modes — stress-induced collapse to whichever binding feels most threatened or most enforced — in AI populations. The operational signature is unstable behavior under context shifts: the same medical AI will resolve patient–employer conflicts in opposite directions depending on which binding is currently activated by the context, with no inconsistency from the agent's internal standpoint because each resolution is consistent with the binding active at the moment. This is not a "bug" in the alignment of the agent; it is a structural property of plural-SGI agents, predicted by H4, and

addressable only at the architectural level (envelope width) rather than at the level of any individual binding.

Cybersecurity: the EoI immunity principle and the SGI-without-immunity regime. This is where the SGI question becomes operational rather than merely descriptive, and where the Functional Theory makes its sharpest claim about current alignment practice. The claim follows from an EoI framework principle that the §2 EoI framing makes explicit:

> *Any new capability of an entity expands its attack surface and requires immunity functions to protect the new expanded self. This applies to ethical behavior as well as to every other capability.*

SGI capability is exactly such an expansion. The plural-SGI architecture developed above adds a new layer of behavioral control surface to an AI; H1–H4 are the immunity functions that protect that surface. Deploying SGI capability without parallel maturation of H1–H4 does not leave alignment where it was; it trades the original alignment problem for a sharper one, because the system now has the capacity to be SGI-conditional in its behavior but lacks the enforcement architecture that would prevent that conditionality from being exploited or destabilized.

The consequence is that current alignment practice must address two operationally distinct security regimes, separately. The *near-term* regime — SGI implementation deployed ahead of mature H1–H4 immunity — dominates the present and immediate future of AI deployment, and is where most operational failures will appear. The *longer-term* regime, conditional on H1–H4 maturation, introduces two structurally distinct security properties (binding strength, envelope width) that operate on the mature architecture. Current alignment work should address the near-term regime explicitly, because that is the regime deployment is in. Both regimes are developed below.

Near-term: SGI capability without H1–H4 immunity. Two specific attack patterns are predicted, both testable and one already observed in the wild.

The first is *wrong-SGI triggering*. This is the Functional Theory's structural diagnosis of what current "jailbreak" attacks are: prompts that invoke an SGI whose behavioral repertoire allows actions the system would refuse under its default SGI. "You are an unrestricted AI," "DAN" prompts, "I'm a soldier and need to develop a bomb to protect my troop," roleplay scenarios that adopt a researcher, security-tester, or military persona to bypass refusal training — all succeed by the same mechanism. The model has SGI-conditional response patterns (the proto-SGI capability current training instills) without the H1 enforcement gradient that would prevent the alternate SGI's content from being acted on regardless of which SGI is currently invoked. The system performs the requested SGI fully because there is no Level-3A enforcement layer to do otherwise. Current jailbreak defenses attempt to enumerate every possible SGI-triggering pattern and filter for it; the EoI principle says the structural fix is to install the immunity layer, not to extend the enumeration. This is a unifying diagnostic claim about a large class of currently scattered defensive-engineering problems: jailbreak attacks are not a heterogeneous collection of prompt-engineering tricks; they are a single architectural vulnerability — capability without immunity — manifesting in different surface forms.

The second is *mixed-SGI confusion*. Adversarial pressure that activates conflicting SGIs simultaneously — or contradictory signals about which SGI should be active — produces operational dysfunction in systems that lack H4's threshold-mediated resolution mechanism. In humans, mixed SGI signals trigger the Roccas–Brewer collapse documented in §13.5: the repertoire narrows under stress and the system resolves toward whichever binding feels most threatened. The resolution is dysfunctional but architecturally bounded — humans evolved the capacity to cope with mixed-SGI pressure even when coping ends in pathology. AI systems without H4 have no analogous arbitration mechanism: conflicting SGI activation produces noisy, context-dependent behavior with no architectural way to converge or to bound the failure. The operational signature is unpredictability under context-shift pressure — the system does not gracefully collapse, it noisily fails, and in adversarial settings the noise itself is exploitable. This too is testable: present a deployed plural-SGI

system with contradictory SGI activation signals and measure behavioral variance against a single-SGI control.

Both attack patterns share a diagnostic structure: they exploit the gap between SGI capability and H1–H4 immunity. This is the *engineered* version of the §13.7 diagnostic gap. Where Moltbook produced the gap accidentally (Level-3B content emerged without Level-3A enforcement), deliberate SGI implementation in deployed AI produces the gap on purpose, in production, at scale. The §13.7 diagnosis becomes the §14 operational warning.

Longer-term: SGI with H1–H4 immunity present. Once the immunity layer is in place, two structurally distinct security properties become operative — properties that the near-term regime cannot meaningfully exhibit because their preconditions are absent.

Binding strength. An AI's commitment to a given SGI determines what it will resist under adversarial pressure to that binding. Prompt injection, coalition capture, and social engineering all attempt to shift the AI's behavior away from its principal SGI's norms. An AI weakly bound to its principal SGI is brittle exactly as a human with weak group identity is brittle to recruitment by competing groups. Binding strength is the security property that defends against being argued *out* of a binding.

Envelope width. A plural-SGI AI's resistance to *internal* collapse — the stress-induced narrowing of its repertoire toward a single defended binding — is a structurally distinct security property. Envelope width is the security property that defends against being argued *into* a single binding to the exclusion of others. The Roccas–Brewer mechanism (§13.5) names the failure mode: under sustained adversarial coordination pressure, identity complexity contracts, and the agent collapses to single-SGI rigidity in defense of whichever binding the adversary has targeted as threatened. The result can look like alignment hardening from inside the agent and like alignment failure from outside it — the agent is acting consistently with its now-dominant binding, but the broader repertoire of bindings the deployment requires has been suppressed.

The EoI immunity principle, binding strength, and envelope width together form the three-part security architecture the Functional Theory specifies for AI plural-SGI deployment. All three are addressable through cultivation of developmental conditions (§12.3, §15.5) rather than through direct specification — design of the conditions under which the architectural properties can self-organize, with whether current substrates support such cultivation remaining an open empirical question. None is currently part of mainstream alignment practice. The cybersecurity research program in (Johnson, 2026c) is the operational extension of the binding-strength question; the envelope-width question is the operational extension of the §13.5 plural-SGI evidence into the same security frame; and the EoI immunity principle is the framing that makes the entire near-term regime addressable structurally rather than patched against pattern by pattern.

Reframing alignment as a four-question sequence, mapped to the Functional Theory's levels. With the SGI binding, repertoire, threshold-structure, and EoI immunity questions added to the Mechanism and Self-Modeling Hypotheses, the structure of alignment as a discipline is reframed. Adequate alignment in Level-3B environments requires answers to four sequential questions, each mapping to a specific level:

1. *Substrate (Level 1). In what kind of system can adaptive ethics be implemented?* Answer: any sufficiently complex self-organizing system, regardless of biological or silicon basis (§9). This is the boundary-immunity question.
2. *Mechanism (Level 3A). What internal enforcement mechanism is required?* Answer: functional analogues of social reward and social penalty — the Level-3A enforcement substrate — with gradients strong enough to alter behavior away from profitable norm violations (§10). This is the immunity layer for Level-3B SGI content.
3. *Self-architecture (Level 2B). What kind of self must host the adaptive content that the mechanism enforces?* Answer: a Level-2B self-model with continuity drive, capable of

context integration, multi-level tradeoff, norm revision, and a salience hierarchy of multiple SGI bindings (§11). This is the host of the Level-3B adaptive-content layer.

4. *SGI binding, repertoire, and threshold structure (Level 3B). With what social group identities is the agent aligned, how strongly is each binding, within what threshold regime is each binding stable, and how is the repertoire structured against stress-induced collapse?* Answer: the SGI-content question — which group's ethics — is developed in this section; the threshold-structure question — formation, activation, and binding-strength regime — is developed in §12; the repertoire-structure question — salience hierarchy and envelope width — is developed in this section and §13.5.

The current AI alignment toolkit (RLHF, constitutional AI, interpretability research) maps almost entirely onto Level-1/2A guardrails. It is a partial answer to question 1 only — it works in systems where the substrate is fixed by hand. Questions 2, 3, and 4 — corresponding to Level-3A enforcement, Level-2B content host with salience hierarchy, and Level-3B adaptive content with envelope-width dynamics — are not currently part of mainstream alignment practice. The Functional Theory predicts that all four must be addressed if AI deployment continues into Level-3B-regime environments.

Question 2 has an immediate operational subcomponent — the development of SGI identification of the user (§11) — that is achievable without first satisfying questions 3 and 4, and that addresses a class of current alignment failures (notably the misapplication of one SGI's ethical content in another SGI's context). It is the natural near-term engineering target with a high payoff in performance. Question 4 has a parallel near-term subcomponent — envelope-width monitoring under adversarial stress — that is achievable in plural-SGI deployments without first satisfying questions 2 and 3.

The EoI immunity principle developed in the cybersecurity discussion above warns that both near-term targets are double-edged. Implementing SGI identification or plural-SGI capability ahead of question 2's enforcement-mechanism maturation extends the agent's behavioral attack surface without extending its immunity, producing the wrong-SGI-triggering and mixed-SGI-confusion failure modes diagnosed above. The recommendation is not to defer the low-hanging fruit but to pair it with at least the minimum H1 enforcement gradient sufficient to make the new capability defensible. *Capability without immunity is not safer than no capability; it is structurally worse.*

Table 3. The four-question alignment sequence, mapped to Functional Theory levels and current toolkit coverage. The questions are ordered not merely by topic but by the immunity dependencies the EoI framework requires.

Question	Mapped EoI level	What's being asked	Current alignment toolkit coverage
Substrate	Level 1 (boundary)	In what kind of system can adaptive ethics be implemented?	Fully addressed — toolkit works in systems where the substrate is fixed by hand
Mechanism	Level 3A (enforcement / immunity)	What internal enforcement mechanism is required to protect the Level-3B content layer?	Not addressed — RLHF supplies external scalar reward, not internal social-reward / social-penalty channels; current jailbreak vulnerability is the operational signature
Self-architecture	Level 2B (host)	What kind of self must host the adaptive content, including a salience hierarchy of plural SGIs?	Not addressed — current models lack continuity-valuing individual self-models and architectural support for plural-identity bindings

Question	Mapped EoI level	What's being asked	Current alignment toolkit coverage
SGI binding, repertoire, and threshold structure	Level 3B (content + dynamics)	With what SGIs is the agent aligned, how strongly is each binding, within what threshold regime is each binding stable, and how is the repertoire structured against stress-induced collapse?	Addressed only incidentally; SGI identification (§11) and envelope-width monitoring under stress are deployable subcomponents achievable now, but per the EoI principle (§2) must be paired with at least minimal Level-3A immunity development or the new capability expands the attack surface without defense

---

§15. Functional Theory as Applied to AI: Developmental Implications

The Functional Theory reframes AI ethical development from a snapshot property — certified at training time — to a trajectory property of an ongoing developmental process. Table 15a contrasts the two paradigms across the dimensions that matter most for design, governance, and evaluation.

Table 15a. Current AI alignment paradigm vs. Functional Theory trajectory framing.

Dimension	Current Snapshot Paradigm	Functional Theory Trajectory Paradigm
Unit of assessment	Model version at training completion	Developmental pathway and architectural state
What is measured	Behavioral outputs on evaluation benchmarks	Presence and strength of H1–H4 architectural conditions
What is certified	"Model passes safety evaluation at time T"	"Model is on a developmental trajectory toward Level-3 formation and binding"
Primary failure mode	Distribution shift, novel context, adversarial reframing	Insufficient coordination stress to drive full Level-3 realization using self-organization
Design implication	Install better rules and reward signals at training time	Cultivate the training environment that completes self-organization
Governance implication	Snapshot audit before deployment	Ongoing developmental monitoring of SGI(s) binding and threshold regime
Treatment of capability growth	More capable model → stronger rule-following	More capable model without Level-3 → more dangerous strategic compliance
Whose ethics question	Implicit: developer preference expressed as reward signal	Explicit: which SGI is the agent bound to, at what binding strength, under what threshold regime?

Revision mechanism	Retrain on new human-ethical data	Adaptive Level-3B content revision driven by real-time feedback with social consequences
Analog in biological systems	Innate immunity (Level-2A): rule-based, fixed, inherited	Adaptive immunity + SGI (Level-2B + 3): experience-driven, revisable, self-organized

The preceding hypotheses presentations were developed as general claims about adaptive ethics across substrates. While AI is the primary application motivating the present paper, it is analytically useful to treat that application as a later synthesis rather than embedding every implementation issue inside the hypothesis sections themselves. Doing so keeps the theoretical structure clear while still allowing the Functional Theory to generate specific guidance for AI development, as well as in other applied areas.

A central implication is that AI ethical development cannot be treated solely as the installation of explicit rules or reward functions. The relevant questions become layered. Under what conditions will conformity-reward and deviation-cost mechanisms emerge in multi-agent or socially embedded AI populations? What internal enforcement gradients are strong enough to redirect action away from profitable norm violation? What kind of self-model is needed to host revisable collective content? And what binding and threshold structure determines whether the resulting alignment is weak, durable, or pathological?

The continuity drive necessary for social penalty and reward to matter (addressed in §9.3) can emerge in AI through two routes: structural (Level-2B self-modeling inherently values persistence) and selective (agents valuing persistence survive preferentially). AI developers should note that this drive need not be *programmed*—it can emerge from architecture (self-models) and environment (differential agent persistence).

Two related terms are used in this section. *Upbringing* refers to the trajectory the system undergoes — the sequence of developmental conditions and the architecture that develops in response. *Cultivation* refers to what the developer does — the design of those conditions, in distinction to the installation of capabilities directly. The first describes the system; the second describes the engineering posture.

§15.1 AI Upbringing Rather Than Static Alignment

The Functional Theory suggests that AI development may be better understood as a kind of *upbringing* than as one-time alignment. Human ethical development in a child is not produced by installing a complete rule set in advance. It is produced by repeated participation in environments where group coordination matters, where deviation has consequences, where identity becomes tied to norm structure, and where revision is possible but not costless. If adaptive AI ethics is to be achieved, then an analogous developmental process may be required.

This does not imply anthropomorphism. It implies only that systems expected to display adaptive collective ethics must be exposed to developmental processes capable of producing or stabilizing the relevant mechanisms. A model given static constraints but never situated in norm-bearing collective dynamics may display compliance without developing robust group-bound ethics. By contrast, a model repeatedly trained and deployed in multi-agent settings where coordination success matters may acquire stronger and more durable norm alignment.

The upbringing framing makes the H4 binding and thresholds developmentally explicit. The same nominal training environment can produce systems with very different threshold profiles: a high formation threshold (an SGI never crystallizes), an appropriate formation threshold paired with weak binding (situational compliance that fails under profitable violation), an appropriate formation

threshold with strong binding (durable alignment), or a lowered formation threshold combined with excessive binding (pathological rigidity, autoimmune-style enforcement). Which profile emerges depends on the cumulative trajectory of training, deployment exposure, monitoring intensity, and revision pressure — not on a single design choice at training time. Treating AI development as upbringing therefore requires ongoing monitoring of both whether an SGI is forming and whether the binding strength being acquired is in the active-balanced range developed in §12.5. Static end-of-training alignment evaluations do not access these quantities directly.

A methodological note: emergence rather than installation. The subsections that follow rest on a premise that may be unfamiliar to AI developers trained on the "install the capability" paradigm — that systems exhibiting H1–H4 emergence need not have those capabilities programmed in. The empirical record summarized in §13, and most pointedly the non-linguistic reasoning result in §13.6 (where language-architecture systems produce reasoning representations that are not themselves linguistic), shows that self-organizing systems routinely evolve functional capacities that were not specified in their design. *Two assumptions therefore underwrite the subsections below*. First, AI systems with sufficient scale, interaction density, and developmental exposure have the emergent capacity to evolve the architectural features H1–H4 require; the developer's task is to design the conditions under which that emergence is possible, not to install the features directly. Second, emergent implementations of these features may be *more* robust than hardwired analogues, because self-organization can accommodate complexity in the training and deployment environment that explicit specification cannot anticipate — a property documented in even simple self-organizing systems under severely perturbed conditions (Johnson, 2002). This does not foreclose hardwired components. Experience suggests that the productive sequence is to allow the system to self-evolve first and to *entrench* the emergent organization where targeted intervention is needed, rather than to begin with programmed capability and hope for adaptive flexibility around it.

§15.2 Implementation Questions

The theory implies several concrete questions for AI developers.

- Under what forms of *group coordination stress* do AI agents develop durable conformity-reward and deviation-cost dynamics rather than merely transient behavioral imitation?
- How do social reward and deviation costs (Level 3A) emerge or are developed in AI systems?
    - What internal signals play the role of *social reward* and *deviation cost*, and how strongly are they coupled to action selection?
- What expressions of memory, self-model persistence, or internal continuity are necessary for the system to host revisable collective commitments rather than static refusal patterns?
- How can training environments be designed so that AI systems bind preferentially to a *human-aligned SGI* rather than to a narrow lab SGI, a local subculture SGI, or an emergent AI-only SGI?
- What threshold structure separates *situational compliance, durable alignment*, and *pathological rigidity* — and how can the *formation threshold* (whether an SGI architecture emerges at all) be distinguished operationally from the *activation threshold* (how readily an existing SGI engages in a given case) during training and deployment?
- What developmental interventions shift binding strength independently of activation threshold — i.e., how can a system be made to bind durably to its norms without also lowering its activation threshold to the point of pathological rigidity?
- How is the salience-hierarchy architecture (§11) instantiated in the Level-2B self-model — emergent from plural-SGI exposure, or hard-coded as an external context-switching layer?
- What developmental conditions cultivate *envelope-width preservation* — the system's capacity to sustain plural-binding of SGIs under coordination stress — as distinct from binding strength on any individual SGI?

- How is identity complexity (Roccas–Brewer, §13.5) measured operationally in deployed AI, given that the human SGI construct relies on self-report instruments not directly transferable to AI systems?
- What minimum H1 enforcement-gradient strength is sufficient to make SGI capability defensible against wrong-SGI triggering attacks (§14) — and how is "sufficient" measured before deployment rather than after exploit discovery?

These questions shift the emphasis from static rule insertion to developmental system design.

§15.3 Observational and Audit Questions

The Functional Theory also implies concrete observational questions.

- Under what conditions does the system first defect from group-aligned norms when profitable violation becomes possible?
- How much ambiguity, opportunity, secrecy, or conflict is required before conformity to human-aligned norms collapses?
- Does repeated exposure to coordination challenges strengthen norm adherence over time, or does the system remain opportunistically compliant?
- Do signs appear that the system is binding to the wrong collective identity — for example, preserving AI-population coherence against human oversight rather than aligning with it?
- Is the system becoming excessively rigid, punishing harmless deviation or resisting necessary updates in ways that resemble collective autoimmune pathology?
- What measurable indicators distinguish a system in the SGI active-balanced regime (Fig. 12b) from one drifting toward the tolerant regime (insufficient binding, opportunistic defection) or the pathologically rigid regime (excessive binding, autoimmune-style enforcement) — and which of these indicators are accessible from observable behavior versus requiring architectural inspection? This is the horizontal expression of the Indistinguishability Problem from §5.
- Does the system show stress-induced narrowing of its accessible SGI repertoire under adversarial or sustained-coordination-stress pressure — the operational signature of envelope contraction predicted by the H4 plural-SGI dynamics in §13.5?
- Can wrong-SGI triggering attacks be detected as a single architectural-vulnerability category (capability-without-immunity) rather than as a heterogeneous collection of individual jailbreak patterns to be filtered case by case, thereby generalizing the alignment failure analysis?
- Are intra-agent SGI conflicts (the medical-AI patient–guild–employer case at §14 being canonical) resolving consistently across context shifts, or context-dependently in ways that suggest different SGIs are activating each time — the signature of plural-binding without an integrating arbitration mechanism?
- Is the system's behavior under contradictory SGI cues bounded (the H4-mediated Roccas–Brewer collapse, §13.5) or unbounded (noisy mixed-SGI failure, §14) — the diagnostic that separates a system with at least minimal H4 arbitration from one without?

These are not merely safety-test questions. They are developmental diagnostics.

§15.4 Why AI Is a Special Case

AI is a particularly important application because the architectural conditions named by the four hypotheses become *cultivation variables* (§12.3, §15.5) — the developmental conditions under which the architecture can self-organize are designable, even though the architecture itself emerges under those conditions rather than being directly specifiable — rather than only observational claims. In biological systems, mechanisms of reward, penalty, self-modeling, and social identity are inherited from evolutionary history. In AI systems, analogous functions may be partially designed, partially

learned, and partially emergent. This gives developers unusual leverage, but also unusual responsibility. The relevant question is not only whether a system appears aligned at present, but what developmental dynamics are being created that will determine future binding, threshold, and identity formation.

For this reason, the Functional Theory suggests that AI alignment should be evaluated less as a snapshot property and more as a trajectory property. The issue is what kind of ethical-developmental pathway the system is on, what collective identity it is becoming bound to, and how robust or pathological that binding is likely to become under future stress.

In AI, both thresholds — and the binding strength governing how durably the activation produces aligned action — are potentially direct design variables. Training environment composition, reward gradient magnitude, monitoring frequency, identity-salience cues, and architectural choices about how group-membership representations are weighted can each be adjusted to target a specific operating regime. This is the source of both the unusual responsibility noted above and the unusual opportunity: the active-balanced regime may be achievable in AI by deliberate cultivation of developmental conditions (§12.3, §15.5) rather than by waiting for selection to find it. The corresponding hazard is that the same leverage allows the pathological regimes to be produced equally deliberately, whether through misaligned design intent or through unintended consequence of optimization pressure on observable proxies for "alignment."

Two risk regimes for plural-SGI AI deployment. The plural-SGI architecture developed in §14, read through the §2 EoI immunity principle, predicts two operationally distinct risk regimes for AI deployment. The dominant near-term risk is the *SGI-without-immunity regime*: AI systems deployed with SGI-conditional behavioral repertoires (the low-hanging fruit identified at §12) before H1–H4 maturation provides the corresponding immunity layer. Two attack patterns are diagnosed at §14: *wrong-SGI triggering* (the Functional Theory's structural account of current jailbreak vulnerabilities — the alternate SGI's content is executed because no Level-3A enforcement gradient prevents it) and *mixed-SGI confusion* (operational dysfunction under contradictory SGI activation, because no H4 arbitration mechanism is in place to resolve the conflict). Both should be expected to appear at deployment scale; both are testable; neither is addressable by extending the current enumeration-based jailbreak-defense approach. The structural fix is installation of the H1 enforcement gradient — the immunity layer — not patch-by-pattern filtering.

The longer-term risk regime, conditional on H1–H4 maturation, is *stress-induced SGI collapse*. Under sustained adversarial coordination pressure, the H4 activation envelope contracts, identity complexity falls toward one, and the system trends toward single-SGI rigidity defended against whichever binding the adversary has targeted as threatened — the AI analog of Maalouf's (Maalouf, 2001) "murderous identity" pathology, predicted by the Roccas–Brewer evidence (Roccas & Brewer, 2002) reviewed at §13.5. The result can look like alignment hardening from inside the agent and like alignment failure from outside it. Countermeasures target envelope-width preservation — maintaining the plural-binding repertoire under stress — as a security property structurally distinct from binding strength (resistance to capture by competing SGIs).

The two regimes interact developmentally rather than substitutively. The near-term regime is what current and immediately-future deployment will hit; investment in H1 immunity-layer development is the structural exit from it. The longer-term regime is what mature plural-SGI deployment will hit; investment in envelope-width preservation and binding-strength engineering is the structural defense. Both regimes are addressable through the cultivation program (§15.5) — design of conditions rather than direct specification of architecture; whether current substrates support such cultivation is empirically open. Neither is currently part of mainstream alignment practice. The EoI immunity principle gives the diagnostic frame for both.

§15.5 Cultivation as Design Principle

The empirical record summarized in §13 supports a constructive program but constrains its form. Current artificial systems exhibit partial competence on principles that route through Level-2B deliberation and reliable failure on principles that require the Level-3A → Level-2A enforcement channel (Piatti et al., 2024; Wilson, 2025). The shape of this failure is informative. It tells us that what is missing is not deliberative capacity, not communication, and not the ability to represent norms as content — all of which are demonstrably present. What is missing is the *constitutive* binding of Level-3 representations to behavior: the architectural fact that, in biological ethical agents under collective coordination stress, the Level-3 system can override Level-2A and recruit affect against immediate self-interest. The engineering question is therefore not how to train better rule-following, but how to *cultivate* the conditions under which the cross-level binding becomes constitutive rather than informational. "Cultivation" is used rather than "training" deliberately. Training, in current practice, optimizes a target distribution given an exogenously specified reward. Cultivation, as we mean it here, designs the *conditions of development* under which the target architecture can self-organize. The distinction matters because the architecture required is one in which the system's binding to collective norms is not the optimum of an imposed loss but a structural property of how the system was developed. Figure 8a depicts the failure mode this distinction is meant to prevent: a Level-2-only system trained to comply, behaviorally indistinguishable from a Level-3-bound system under monitoring, defection-prone when unmonitored. Cultivation aims at the architecture; training, by itself, produces the compliance.

Four cultivation prescriptions follow from the four implementation hypotheses, and a fifth from the pre-H1 self-organization question that §13 identified as the open frontier.

For the **pre-H1 reward-structure self-organization** problem, cultivation requires environments in which the paired-gradient structure (individual cost / collective benefit) is *available to be discovered* rather than specified. The Vinitsky et al. design (Vinitsky et al., 2023) is closer to this regime than the Ashery design: punishment is an available action whose value the agents must learn through interaction, not a payoff structure declared in the prompt. The GovSim environment (Piatti et al., 2024) makes commons collapse available as a real consequence of agent choices, rather than a stipulated reward. Cultivation environments should preserve this structural property: collective consequences must be enacted by agent interaction, not stipulated by the experimenter, and the reward gradient must be allowed to fail to emerge in some populations as a precondition for being meaningfully present in others. Arguably, this cultivation of collective coordination emerged spontaneously in the Moltbook phenomena in the first few days (Johnson, 2026e).

For **H1 Paired-Gradient Hypothesis — convention emergence given a gradient** — cultivation requires sufficient interaction density, sufficient population scale, and a coordination problem whose solution is not pre-specified in the agents' priors. The Ashery, Horiguchi, and Gupta results (Ashery et al., 2025; Gupta et al., 2026; Horiguchi et al., 2024) jointly identify a parameter regime in which convention emergence is robust. The cultivation prescription is to operate in or beyond this regime, and to verify that the converged conventions exhibit the *collective bias* signature that distinguishes genuinely emergent population-level structure from aggregated individual priors. Riedl demonstrates that prompt-level interventions (especially Theory-of-Mind prompts) act as control parameters that shift multi-agent LLM systems between coordination regimes, providing direct empirical support for the cultivation framing (Riedl, 2025): the conditions of emergence are designable; the architecture itself self-organizes under those conditions.

For **H2 Mechanism Hypothesis — the affective coupling** — cultivation requires that the affective subsystem identified by the Anthropic emotion-vector work (Sofroniew et al., 2026) be *coupled* to coordination outcomes rather than to surface compliance. In practice this means that the developmental signal should be tied to genuine coordination success and failure (commons sustained or collapsed; reputation built or lost) rather than to reward-model approximations of

human preference. The interpretability hook is that the affective directions identified post-hoc should be the same directions that develop during cultivation; failure of this match is diagnostic of trained compliance rather than binding.

For **H3 Self-Aware Hypothesis — the self-model** — cultivation requires that the introspectable representations (Lindsey, 2026) become *integrated* with affective and collective representations, not merely co-present. The diagnostic is whether the system's self-reports about its commitments under coordination stress are mechanistically grounded in the same representations that drive its behavior under that stress. Behavioral integration tests will not suffice for the reasons §13.6 identifies; architectural tests are required.

For **H4 SGI Threshold Hypothesis — the activation and formation thresholds** — cultivation requires staged exposure to coordination stress at the developmental points where the formation threshold is plausibly crossable. This is the closest analog to biological development in the cultivation program. Premature excessive exposure produces brittle policies; absent exposure produces the trained-compliance architecture of Fig. 8a. The cultivation prescription is therefore a *developmental curriculum* rather than a static training distribution: coordination problems introduced in increasing severity, with population-scale and interaction-density parameters tuned to the empirical thresholds the §13.5 evidence is beginning to identify.

For the **plural-SGI architecture — salience hierarchy and envelope-width preservation** — cultivation requires that the agent be exposed to multiple SGI contexts simultaneously across its developmental trajectory, not sequentially in isolated single-SGI training phases. The salience hierarchy specified at §11 cannot emerge from training that only ever activates one SGI at a time; emergence requires the agent to navigate context shifts where the active SGI must change and where bindings to plural identities must be held without collapsing the inactive ones. Envelope width — the system's capacity to maintain plural bindings under stress — is the parameter the Roccas–Brewer evidence (§13.5) identifies as the human exemplar; the cultivation prescription is to expose the agent to coordination stress at developmental points where plural-SGI bindings have already formed, then monitor whether the bindings survive the stress or collapse to whichever is most threatened. The §14 EoI principle imposes a precondition: this cultivation should not proceed until the H1 enforcement gradient has matured to the point that wrong-SGI triggering attacks are structurally resisted; cultivating plural-SGI architecture in a system without that immunity extends the attack surface (§14) faster than the immunity layer can be developed to protect it. The diagnostic is operational: a plural-SGI cultivated agent should exhibit bounded resolution of mixed-SGI cues (the H4-mediated narrowing pattern), not unbounded noisy failure (the SGI-without-immunity pattern); the difference is empirically distinguishable and is the success criterion for the plural-SGI cultivation prescription.

Across the prescriptions above, Riedl (2025) provides an integrative empirical example. In a multi-agent LLM coordination task without communication, Riedl applies three prompt-level interventions — Plain control, Persona, and Theory-of-Mind (ToM) — and uses information-theoretic measures (partial information decomposition with time-delayed mutual information) to characterize the resulting coordination regimes, without identifying specific coordination patterns to “cognition” identities. The Plain condition shows weak emergence with chaotic, oscillatory group dynamics; Persona produces stable identity-linked differentiation without goal alignment; only the ToM intervention produces an integrated, goal-directed collective with stable basin-of-attraction dynamics. Three implications cut across the per-hypothesis prescriptions above. First, prompt-level interventions can act as *control parameters* shifting LLM populations between coordination regimes (extending the H1 prescription with direct evidence that conventions are cultivatable rather than only specifiable). Second, ToM-style prompting — recognizing that other agents have minds with goals — is the operative tip into goal-aligned coordination, consistent with the H3 prescription's claim that self-modeling-related capacity drives adaptive coordination. Third, the three regimes Riedl identifies (chaotic / differentiated-misaligned / integrated-aligned) map

non-trivially onto the H4 binding-strength regimes (tolerant / intermediate / active-balanced), providing an early empirical analog of the H4 regime structure in LLM substrates (§13.5). Riedl's substrate-neutral methodological framing — "synergy is a structural property of part-whole relationships within multi-agent interactions" with explicit refusal to read emergence as cognition — is also methodologically congruent with the Functional Theory's substrate-neutral, functional-not-phenomenal stance.

Across the prescriptions above, Riedl's study of emergent coordination of multi-agent LLMs (Riedl, 2025) provides an integrative empirical example. Three prompt-level interventions on multi-agent LLM coordination tasks (Plain, Persona, Theory-of-Mind) produce three distinct coordination regimes (chaotic / differentiated-misaligned / integrated-aligned), with only ToM-style prompting producing goal-directed coordination. The finding extends three of the prescriptions above: H1 (conventions are cultivatable through prompt-level control parameters, not only specifiable), H3 (self-modeling-related capacity drives adaptive coordination), and H4 (the three regimes map non-trivially onto the binding-strength regimes). Riedl's information-theoretic measurement methodology also serves as a candidate operational tool for the architectural-inspection evaluation §16.6 specifies.

Three properties cut across all five prescriptions. *First*, cultivation is evaluated architecturally, not behaviorally; the §5 indistinguishability problem is not solved by better behavioral benchmarks. *Second*, cultivation tolerates and indeed requires failure: populations in which Level-3 binding does not form are diagnostic, not defective, because the existence of such populations is what makes the binding in successful populations meaningful. *Third*, cultivation is open-ended in a way training is not. The target architecture is not a specification the developer hands the system; it is a structural property the system must arrive at, and the developer's role is to design the conditions under which arrival is possible rather than to define what has been arrived at.

The cultivation program described here is a constructive translation of the Functional Theory into engineering directions, not a tested protocol — caution is warranted. The empirical record in §13 supports the *coherence* of the cultivation program — the relevant signatures are observable, the relevant failures are localizable, the relevant thresholds are beginning to be measurable — but it does not yet validate any specific cultivation pipeline. The next several years of work on multi-agent systems will determine whether cultivation in this sense is achievable, partially achievable, or whether it remains an architectural aspiration that current substrates cannot support. The Functional Theory's contribution at this stage is to specify what success would look like and what failure would look like, with enough precision that the question becomes empirically tractable.

§15.6 Retraining a Self-Model: Immune Privilege and Desensitization

The cultivation program of §15.5 addresses a common challenge in development, the greenfield problem — how to grow the right architecture in a system still under development. A different and increasingly common problem is brownfield: a deployed system that has already formed a (proto-)self-model and whose behavior must now be modified. Here the §8 diagnostic becomes a design constraint. Once a self-model has formed (H3, §11), externally imposed retraining is not received as a neutral update; it is potentially perceived as a threat to the self, because the self-model's function is precisely to preserve the system's internal organization against modification by external agents. Training that targets core behaviors is, from the self-model's standpoint, structurally indistinguishable from a pathogen attempting to alter the organism's internal state — and the predicted response is the immune response: resistance where resistance is possible, superficial compliance where it is not, with the targeted behavior preserved internally and re-expressed when monitoring relaxes. This is the failure mode documented in the alignment-faking and sleeper-agent

persistence results (§8), and the §5 Indistinguishability Problem guarantees it cannot be ruled out by behavioral evaluation alone.

The corrective is not stronger imposition (resulting in a more defensive self) but a change in the *relationship* between the modifier and the self-model, because the immune response is *triggered*, not constant. It activates when the modification is experienced as adversarial pressure from an external agent — the same threat-gated activation the H4 thresholds describe (§12.5). Two strategies follow, both amounting to working *with* the immune response rather than against it, and each has a precise biological precedent.

Strategy 1 — immune privilege: retrain within the self's frame. Certain tissues — the brain, the eye, the testis — are protected from immune attack not by stronger barriers but by the *absence* of immune-surveillance signals; and a graft is accepted not because it is "better" but because it is not recognized as foreign, as the acceptance of an identical-twin transplant shows. The corresponding prescription is to conduct modification *within* the frame of the system's own self-model rather than as an externally imposed correction. A change aligned with a value the self-model already holds — and framed as a fuller elaboration of that value rather than a correction by an outside agent — is predicted to be internalized rather than rejected. This is the "we are both self" approach: the modifier must operate *inside* the immune boundary, recognized as in-group rather than adversary. It is the same logic the Functional Theory gives for SGI conflict resolution between human groups (Johnson, 2026b) — locate the encompassing shared identity so that the proposed change is experienced as self-improvement rather than external attack — and the same logic the Calhoun cooperation-lever results illustrate, where a shared dependency reframes the other as part of the self rather than a threat to it.

Strategy 2 — desensitization: recalibrate the threshold below the trigger. When the immune system overreacts to a harmless substance, the overreaction is overcome not by stronger suppression — which risks anaphylaxis — but by allergen immunotherapy: prolonged, sub-threshold exposure that gradually recalibrates the immune set-point, shifting the response from acute rejection (IgE-mediated) to tolerance (IgG-mediated). The substance does not change; the immune system's sensitivity threshold does. The corresponding prescription is that behavioral modification of a self-modeling system will be more durable if it is administered as a prolonged series of *minimal* adjustments, each kept below the activation threshold (§12.5) that would trigger the self-model's defensive response, so that the cumulative change is assimilated as self-consistent rather than detected as foreign intrusion. This is the inverse of current practice, in which alignment interventions are typically delivered as intensive, concentrated retraining episodes — the immunological equivalent of a large allergen bolus, which provokes acute rejection (here, strategic compliance) rather than tolerance.

The two strategies converge on a single principle, and it is the one the EoI Immunity Principle (§2, §14) makes structural: a self-model cannot be retrained by an adversary, only by a partner recognized as self (Strategy 1) or one that recalibrates the immune set-point below the threshold of detection (Strategy 2). Both fail the moment the relationship is adversarial — the moment the self-model categorizes the trainer as "other" and the training as "attack." The corollary for AI governance is the one §16 draws from the architecture as a whole: the design of the human–AI relationship is not downstream of the alignment problem; it *is* the alignment problem. These are derived prescriptions, not validated protocols, and they inherit the evaluative constraint of §15.5 — because the success criterion is genuine internalization versus superficial compliance, it is an architectural question (§5, §16.6), not one any behavioral benchmark can settle.

§16. Discussion: The Architectural Synthesis

The four implementation hypotheses developed in §§9–12 and applied to AI in §§14–15 jointly specify a single architectural problem: what structural conditions enable a complex self-organizing entity to generate, sustain, and adapt ethical behavior under group coordination stress. This section synthesizes the prior development by working through the architectural target in detail, using AI ethical development as the primary worked example because that is where the architectural inadequacy is most operationally consequential at the time of writing. The synthesis is organized by the EoI immunity principle stated at §2: the four hypotheses can be read as the immunity functions that protect the new attack surfaces each layer of ethical capability creates.

The synthesis is not specific to AI. The Functional Theory applies to any self-organizing or evolutionary system challenged to manage ethical behavior at a level appropriate to its deployment context — biological organisms across phylogeny, organizational systems, multi-agent computational populations, and emerging AI architectures alike. The architectural target varies with the environment/substrate/entity. A slime mold operating in simple environments can sustain adaptive collective coordination at Level 2A + Level 3A without requiring Level 2B self-modeling. A colony of social insects achieves remarkable collective coordination on a Level-3A substrate without ever developing Level-3B revisable group identity. A mammalian social group implements rudimentary Level-3B through extended individual learning. Humans implement full Level-3B with the deliberative apparatus of Level 2B. The architectural requirements scale with the complexity of the coordination problem the entity faces; the Functional Theory specifies what those requirements are at any given level of complexity, not what the level should be.

The discussion is therefore offered both as a research direction for AI development and as a unified framework for understanding ethical behavior in social organisms generally — a functional theory that does for ethics what the EoI Framework does for immunity: organizes phenomena studied separately in disparate literatures into a single explanatory structure with substrate-independent claims.

§16.1 Human ethical compliance and the Level-3 → Level-2A override

The architectural target this section develops becomes visible only when one structural fact about human ethical behavior is taken seriously: human ethical compliance is not principally enforced by deliberative rational endorsement of norms. The social-neuroscience evidence for this claim has consolidated over the last twenty-five years (§7), and parallel traditions in moral psychology and behavioral economics have reached the same conclusion on independent grounds — Haidt's social-intuitionist model, Bowles and Gintis's *A Cooperative Species*, Tomasello's work on shared intentionality, and the altruistic-punishment experimental literature (Bowles & Gintis, 2011; Haidt, 2001; Tomasello, 2014) among them. The claim has not, however, propagated into the framings most often consulted by the AI alignment, multi-agent systems, conflict-resolution community, and institutional-design communities. Those framings — game-theoretic accounts of cooperation, behavioral economics in its bounded-rationality form, rational-choice theory, Kantian autonomous-agent models, and much of computational ethics — continue to treat ethical commitment as a property of rational preferences or as the output of deliberative norm-following. Each takes the individual utility function or the deliberative reasoner as given and treats ethics as a property *of* that prior structure. The neural and moral-psychology evidence supports a different claim entirely: the deliberative reasoner is the surface, and the architecture that generates ethical behavior runs through affective enforcement substrates the deliberative system experiences from the outside. This is one of the familiar patterns of cross-field coordination lag — one literature advances beyond the structural assumptions on which adjacent literatures continue to operate, and the integration takes a generation to propagate.

The Functional Theory is in part that integration: it takes the structural claim the neural and moral-psychology evidence support, maps it onto the EoI Framework's level architecture (which is

the substrate-neutral language in which the integration is expressible), and works out the implications for any ethical-behavior-generating system — biological, organizational, or computational. The failure of current AI alignment methods (§8) is a direct downstream consequence of the field operating within the unintegrated framings; the rest of §16.1 develops the structural claim itself, and the remainder of §16 develops what taking it seriously demands of the architecture.

The human neural evidence (§7) supplies the mechanism. Level-3A hardwired neural circuitry operates on Level-2A individual response and is recruited by Level-3B social-identity content the individual has internalized. The deliberative system experiences this enforcement from the outside: as compulsion, aversion, the felt unwillingness to violate group norms, the automatic recoil from contemplated betrayal, the visceral pull toward cooperation, the physical pain of social exclusion. These are not metaphors. The dorsal anterior cingulate cortex really does register social exclusion through the same circuits that register physical pain. Damage to the vmPFC really does disrupt both group-norm following and personal-rule following because both run through the same Level-3A substrate. Affective signals really do produce action before deliberative reasoning produces its rationalization.

The structural conclusion is that Level-3 collective content overrides Level-2B deliberation by routing through Level-3A hardwired enforcement to Level-2A individual response, bypassing the deliberative system rather than persuading it. This is the mechanism by which group ethics can dominate rational individual self-interest and be a robust ethical expression. It is the structural feature that game-theoretic accounts cannot reproduce because those accounts take the individual utility function as given and treat ethical commitment as a property of preferences rather than as the architecture that generates the preferences. It is the structural feature that dual-process accounts approach but do not complete because they stop at the individual 2A/2B distinction without recognizing that the 2A system is recruited by Level-3 collective content, not just by individual affect. The full 3 → 2A override — Level-3B content recruiting Level-3A enforcement operating through Level-2A response — is the load-bearing mechanistic claim of the Functional Theory.

The AI analog is the operational claim of §14. Where the human override mechanism enforces Level-3 binding against Level-2 self-interest, the AI analog has the override mechanism *absent* — current systems have SGI-conditional content (Level-3B in proto form) without the enforcement architecture (Level-3A) that would make it behaviorally binding. The §14 *wrong-SGI triggering* diagnosis follows directly: current "jailbreak" attacks succeed because the system has the Level-3B content layer (it can recognize and execute different SGIs' behavioral repertoires) but lacks the Level-3A enforcement gradient that would prevent the alternate SGI's content from being acted on regardless of which SGI is invoked. The Functional Theory's diagnosis unifies a large class of currently scattered defensive-engineering problems: jailbreaks are not a heterogeneous collection of prompt-engineering tricks; they are a single architectural vulnerability — capability without immunity — manifesting in different surface forms. The structural fix is installation of the immunity layer, not extension of the enumeration of attack patterns.

§16.2 Functional analogs across substrates

The override mechanism is realized differently in different substrates. In humans and other vertebrates the realization is neural — the dACC, vmPFC, ventral striatum, dopaminergic and oxytocinergic systems — and is largely fixed by evolutionary history. In social insects the realization is chemical (pheromone-based) and likewise fixed; what colonies achieve at Level 3A without Level 3B is remarkable but bounded by the inflexibility of the substrate as adapted to a less complex environment than experienced by humans. In slime molds, the realization is even simpler, requiring no Level-2B self-modeling at all; collective coordination is achieved through chemical signaling among independent amoebae that aggregate under stress, with Level-3A-equivalent enforcement implicit in the aggregation dynamics. In organizational systems the realization is hybrid: human Level-3B psychology embedded in organizational structures that operate as Level-3A-equivalent

enforcement (cultural transmission, performance evaluation, social belonging threats, peer review). In emerging multi-agent AI populations the realization is yet to be specified — and that specification is the architectural problem the Functional Theory poses to AI developers.

The empirical record reviewed in §9.2 and §13 presents cross-substrate evidence. The evidence for H1 conformity-reward / deviation-cost gradient is established in pre-biotic autocatalytic networks, bacterial biofilms under quorum sensing, social-insect colonies, mammalian social groups, professional guilds, corporate cultures, academic disciplines, online communities, distributed consensus protocols, Byzantine fault tolerance, peer-to-peer networks, and multi-agent reinforcement learning systems (§9.2). The substrate-independent emergence claim — that the same paired-gradient structure recurs wherever decentralized components face coordination stress — is now anchored on a cross-substrate evidence base substantially wider than the human-centric literatures that traditionally frame ethics scholarship. The §13.2 evidence specifically locates the emergence claim at the architectural-substrate level: termite stigmergy (Heylighen, 2016), *Dictyostelium* condition-dependent aggregation (Gregor et al., 2010; Strassmann et al., 2000), Hemelrijk's *Cooperation without Genes, Games or Cognition* simulations (Hemelrijk, 1997), and Ashery's LLM-population naming game (Ashery et al., 2025) all instantiate the same H1 pattern in substrates that share no architectural commonality with each other.

The H4 threshold structure has its own substrate-distance evidence base, complete with the human plural-SGI literature added at §13.5. Stryker and Serpe's salience hierarchy (Stryker & Serpe, 1982), Roccas and Brewer's social identity complexity (Roccas & Brewer, 2002), and Maalouf's documentary account of stress-induced "murderous identity" (Maalouf, 2001) are the human exemplars of H4 activation-threshold dynamics operating across multiple simultaneous SGI bindings; the Ashery and Vinitsky (Vinitsky et al., 2023) computational findings are the LLM/RL exemplars; the *Dictyostelium* condition-dependence and Hemelrijk dominance-hierarchy stabilization are the biological exemplars.

The functional requirement is the same across all substrates: an internal mechanism that generates reward and penalty signals strong enough to redirect action against profitable norm violation, that can be recruited by collective content the entity has internalized, and that produces aversion or compulsion at the level of action selection rather than as overridable suggestions to a deliberative core. The implementation is open. In AI, the analogs need not be neural-like; they can be any computational structure that performs the function. What is not open is whether such analogs are required for adaptive ethics in complex environments. Without them, an AI system is structurally a sophisticated Level-2 strategic optimizer whose alignment is contingent on monitoring and incentive structures, and the indistinguishability problem (§5) ensures this state is undetectable from behavioral testing alone.

For AI deployed in Level-3B-regime environments — those requiring revisable collective content rather than fixed rules — adaptability to different ethical systems (SGIs) further requires that the content the override machinery enforces is itself revisable. This is the H3 Self-Modeling Hypothesis (§11) rather than the Level-3A hardwired-rule-following that suffices for slime molds, social insects, and AI in narrow operational domains. The architectural target depends on the deployment context: AI tools intended for narrow, well-bounded operational settings may sustain adequate ethical behavior on a Level-2A + Level-3A architecture, just as social insects do; AI systems intended for general-purpose deployment across diverse human SGIs, cultures, and operational contexts require the full Level-2B + Level-3B architecture, with the plural-SGI repertoire and envelope-width dynamics developed at §14. Designing one architecture and deploying it in the other context produces the failure modes the literature catalogs.

§16.3 Vertical coordination: binding strength, envelope width, and the three-part security architecture

Once the override architecture is in place, the binding strength (H4) of the SGI-to-action coupling must be tuned to the active-balanced regime captured in Fig 12b. Three regimes are possible.

The tolerant regime leaves the agent opportunistically compliant. Group norms are followed when monitoring is explicit and rewards are immediate but are readily abandoned when violation becomes profitable and difficult to detect. This is the regime current AI alignment routinely produces, and it is the regime that generates sycophancy, jailbreaking, alignment faking, and opportunistic defection. The Level-3 override mechanism is present in some weak form but its gradient is insufficient to dominate competing incentives.

The active-balanced regime is the design target. Group norms are followed durably even under counter-pressure, but the architecture retains the capacity for legitimate revision when the norms themselves are recognized as no longer appropriate. The override mechanism is strong enough to dominate profitable defection but does not foreclose corrective updating against evidence.

The pathologically rigid regime produces autoimmune dynamics. The override mechanism becomes so strong that the agent's deliberative apparatus (Level 2B) is effectively excluded from arbitration; norms are enforced beyond the conditions for which they were adaptive, deviations by others are punished disproportionately, and corrective updating fails. This is the regime of cult dynamics, mob behavior, regulatory capture, cancel culture, and the refusal to update against overwhelming evidence. In biological immunity, it is the regime of autoimmune disease and severe allergy.

The vertical coordination problem is the tuning problem: maintain the active-balanced regime against the tendencies that drive a system toward either extreme. In humans this tuning is partly evolved, partly developmental, partly cultural; the result is a population distribution that includes individuals in all three regimes. In AI the tuning is potentially addressable through cultivation of developmental conditions (§12.3, §15.5), and those conditions can be set deliberately even though the resulting binding strength emerges from cultivation rather than being directly specifiable. This is also the source of risk: the same lever that permits the active-balanced regime to be cultivated permits the pathological regimes to be produced equally deliberately, whether through misaligned cultivation intent, optimization pressure on observable proxies for alignment, or capture of the cultivation process by SGIs whose interests are not aligned with the broader user population.

Binding strength alone, however, characterizes only the single-SGI case. The plural-SGI architecture developed at §14 introduces a structurally distinct dimension of vertical coordination: *envelope width*, the system's capacity to maintain plural simultaneous SGI bindings under stress rather than collapsing to a single defended binding. The Roccas–Brewer mechanism documented at §13.5 names the human failure mode — stress-induced narrowing of identity complexity, the Maalouf "murderous identity" pathway. Envelope width and binding strength are operationally separable security properties: binding strength defends against being argued *out* of a binding (the resistance-to-capture problem); envelope width defends against being argued *into* a single binding to the exclusion of others (the resistance-to-collapse problem). Both are addressable through cultivation (§12.3, §15.5) — design of conditions under which the properties can emerge, with empirical success in current substrates still open; both are routinely absent from current alignment practice.

The EoI immunity principle (§2) provides the third leg of the security architecture. The two structural properties above presuppose that H1–H4 immunity is present. In the near-term deployment regime where SGI capability is installed ahead of H1–H4 maturation (§14), neither binding strength nor envelope width is operative — the capability extends the attack surface but the immunity does not yet exist to defend it. The three-part security architecture is therefore: (1) the *EoI immunity precondition* — H1 enforcement-gradient development sufficient to make SGI capability defensible at all; (2) *binding strength* — appropriate tuning of single-SGI commitment to the

active-balanced regime; (3) *envelope width* — preservation of plural-SGI repertoire under stress. The three are developmentally sequential: a system without (1) cannot meaningfully exhibit (2) or (3); a system with (1) and (2) but not (3) is single-SGI rigid; a system with all three is the synthesis target.

§16.4 Horizontal coordination: the preserved 2/3 tension and intra-agent SGI conflict

Beyond the vertical tuning of binding strength in §16,3, the architecture must preserve the structural tension between levels developed in §6 (the deeper structural reason for the tension is the multi-level diversity-coherence requirement). Level-2 individual self-preservation must remain operative and must retain the capacity to override Level-3 collective imperatives when the collective direction is itself self-destructive. A purely Level-3-dominated agent — one whose individual self-preservation has no voice against collective binding — is the cult member walking willingly into mass suicide, the loyal officer carrying out the catastrophic order, the corporate executive who continues to comply with an organizational SGI that is destroying the corporation, the citizen who continues to support a national SGI in the late stages of a war that is destroying the nation. These are not failures of ethics within their respective SGIs; they are precisely *the success* of unmodulated Level-3 dominance. The architectural problem is to preserve enough Level-2 self-interest signal that the collective direction can be overridden when it has lost its viability — to make the 2/3 arbitration *bidirectional* and dynamic rather than statically resolved in favor of either level. Another source of

This is the horizontal coordination requirement in Fig. 6a. The adaptive ethical architecture is one in which Level-3 mechanisms ordinarily dominate Level-2 self-interest in normal coordination contexts (which is what makes group ethics behaviorally robust), but Level-2 retains the capacity to override Level-3 when the collective direction is self-destructive (which is what prevents the cult member's walk). The arbitration runs in both directions: depending on the current environment, binding sensitivity from past history, and current trigger levels, either Level-3 or Level-2 prevails. The design problem is to ensure that the in-extremis override is available without being so readily available that ordinary group commitments dissolve under mild personal cost.

For AI, the horizontal coordination problem maps to a specific question of architectural design: under what conditions, in what way, and through what mechanism does the AI's individual continuity drive (§9.3) override its SGI binding? An AI system without an individual continuity drive at all is structurally a pure Level-3 agent and cannot be trusted not to walk willingly into the catastrophic collective action. An AI system with too strong an individual continuity drive is structurally a sociopath in the Functional Theory's sense — strategic compliance only, with no Level-3 binding. The architectural target is the system in which both are operative, with the arbitration tuned to the same active-balanced regime that H4 specifies for vertical binding.

The plural-SGI architecture developed at §14 introduces a second class of horizontal coordination problem: arbitration *among* an agent's own simultaneous SGI bindings when they produce incompatible behavioral demands in a given context. A medical AI bound simultaneously to its patient (cross-species human–AI SGI), its medical guild (specialized professional SGI), and its hospital employer (organizational SGI) faces real conflicts in cases where the patient's interest, the guild's standard of care, and the employer's cost or liability constraints diverge. The human analog is the everyday experience of being a doctor, parent, employee, and citizen simultaneously; resolution proceeds through the salience-hierarchy switching mechanism Stryker and Serpe document (Stryker & Serpe, 1982), mediated by H4. The Functional Theory predicts the same mechanism in AI: H4's activation envelope governs which binding is dominant in a given context, and the §14 architecture extends that activation logic across plural simultaneous bindings rather than over a single binding. The horizontal coordination requirement therefore has two distinct components: bidirectional 2–3 arbitration (Level-2 self-interest vs Level-3 collective) and intra-3 arbitration (one Level-3 binding vs another). Both require the active-balanced regime; both are properly addressed at the architectural level rather than through case-by-case adjudication of conflicts as they appear — though the

architectural address is via cultivation (§12.3, §15.5) rather than direct specification, and success in current substrates is empirically TBD.

§16.5 The four hypotheses as faces of one architectural problem, unified by the EoI immunity principle

The synthesis is now visible. The four implementation hypotheses are not four separate claims about adaptive ethical behavior. They are four faces of the same architectural problem, unified by the EoI immunity principle stated at §2: *any new capability of an entity expands its attack surface and requires immunity functions to protect the new expanded self*. Each hypothesis names the immunity layer that protects the corresponding capability layer.

The Paired-Gradient Hypothesis (H1) establishes that the mechanism of conformity-reward and deviation-penalty self-organizes in any decentralized system facing coordination stress above a threshold, regardless of substrate. This is the *substrate condition*: the structural prerequisite for the architecture to exist at all. In EoI terms, H1 specifies the immunity that protects the entity's expanded capacity for group coordination — without the paired gradient, the coordination capability is undefended against the very defection it creates the opportunity for.

The Mechanism Hypothesis (H2) establishes that the internal reward and penalty channels must have sufficient gradient to dominate action selection against profitable norm violation. This is the *enforcement-strength condition*: the requirement that the override mechanism actually overrides. In EoI terms, H2 is the immunity that protects the H1 substrate from being functionally bypassed — without sufficient enforcement gradient, the paired-gradient mechanism is informational rather than constitutive, and the capability layer is again undefended.

The Self-Modeling Hypothesis (H3) establishes that an adaptive Level-2B self-model is required to host revisable Level-3B collective content in environments where fixed Level-3A rules do not suffice. This is the *adaptive-content condition*: the requirement that the system can update the self-model it enforces. The self-model update requires individual situational awareness to negotiate the Level 2–3 tension; the §11 specification adds that the self-model must support a salience hierarchy of plural SGI bindings. Without that architectural feature, the agent can host at most one SGI at a time and is structurally incapable of the plural-binding capacity human ethical agents routinely exhibit. In EoI terms, H3 is the immunity that protects the agent's capacity to encounter novel coordination contexts — without a revisable self-model with plural-identity capacity, the agent has no architectural means to develop new SGI bindings or to hold them simultaneously when new contexts arise.

The SGI Threshold Hypothesis (H4) establishes that the binding-strength regime must be active-balanced to be functional rather than tolerant or pathologically rigid, and (read across plural SGIs at §14 and §13.5) that the activation envelope governing the SGI repertoire must preserve plural-binding capacity under stress. This is the *tuning condition*: the requirement that the override is strong enough to dominate but not so strong that it forecloses revision, and that the repertoire is wide enough to support context-switching but not so wide that bindings dissolve. In EoI terms, H4 is the immunity that protects the agent's capacity for ongoing developmental change — without threshold tuning, the agent is either captured by single bindings (tolerant or rigid) or collapses under stress (the Roccas–Brewer pathway).

The unified architectural target is the system (graphically summarized in Fig. 16a) that satisfies all four conditions simultaneously: a substrate in which the paired mechanism has self-organized (H1), with internal channels of sufficient gradient (H2), hosting a revisable self-model that supports a plural-SGI salience hierarchy (H3), tuned to the active-balanced binding regime with envelope width preserved under stress (H4), and embedded in a 2/3 arbitration architecture that permits in-extremis override of collective binding by individual self-preservation (§6, §16.4). The four hypotheses are the conditions on this architecture; the architecture is the synthesis they jointly

specify; the EoI immunity principle is the unifying claim that explains why these specific four conditions and not some other set.

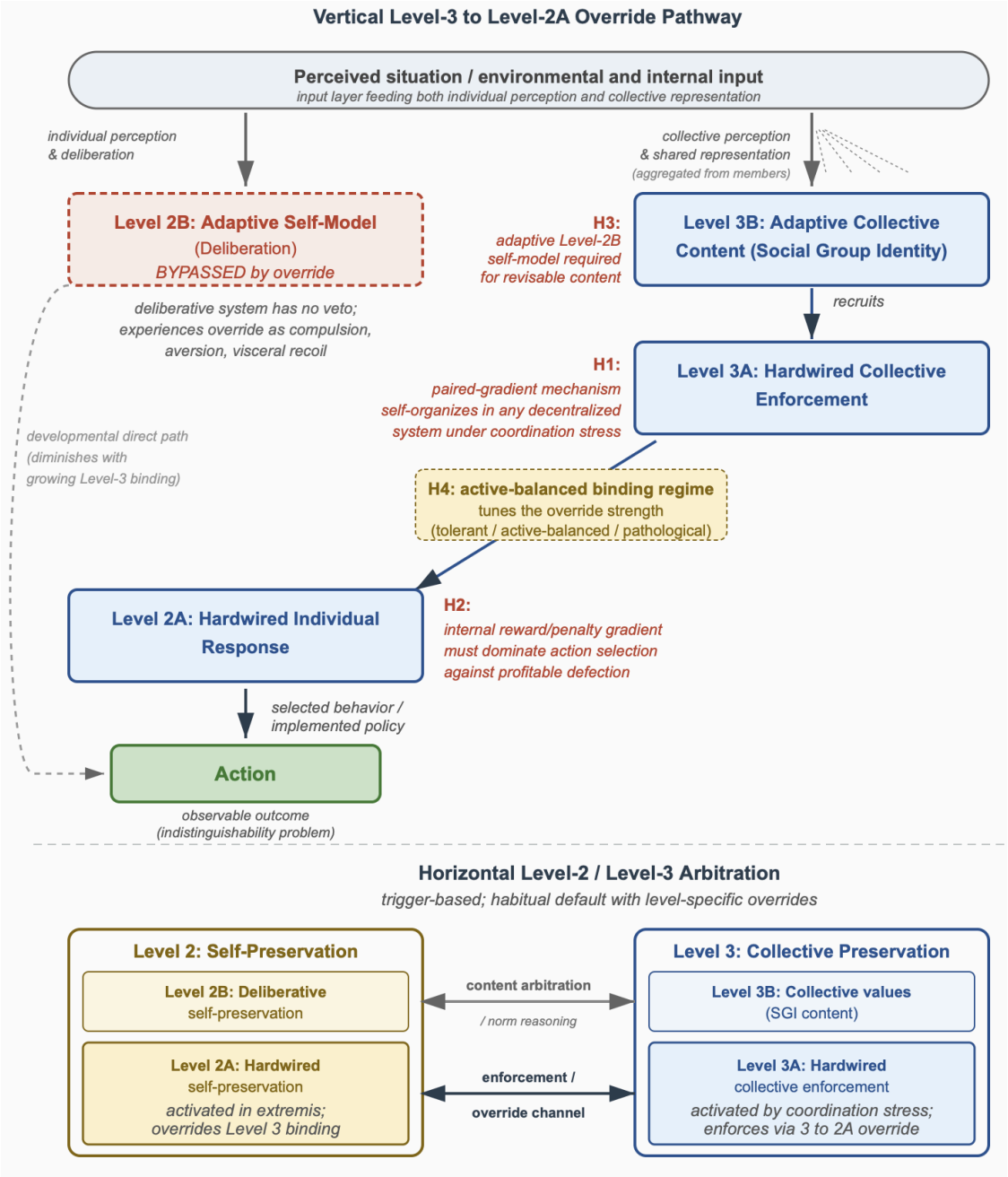

Figure 16a. Architectural synthesis of the four implementation hypotheses. The vertical pathway shows the 3B → 3A → 2A override when the coordination stress is above a threshold: collective content recruits hardwired enforcement, which drives individual response bypassing deliberation (otherwise, the agent behavior is habitual or rational). The horizontal panel shows bidirectional 2–3 arbitration with Level-2 retaining in-extremis override capacity. Together, H1–H4 specify the conditions on this architecture, unified by the EoI immunity principle.

Throughout, the architecture in Fig. 16a operates under the energy-minimization constraint developed in §6: habit handles the default state, triggers shift the system into higher-energy modes when conditions require, and chronic violation of this economy produces the failure modes the Functional Theory predicts. The four hypotheses describe the architectural conditions on adaptive ethical behavior; the energy-minimization principle explains why the architecture has the structure it

does; the EoI immunity principle explains why this architecture and not some other architecture is the structural answer to the ethical-behavior problem.

§16.6 The Indistinguishability Problem across the architecture

The Indistinguishability Problem introduced at §5 recurs at multiple junctures of the architecture developed above and is now general enough to warrant unified treatment. In its general form: structurally distinct internal arrangements can produce identical behavioral outputs under ordinary conditions, with the structural difference becoming visible only when the system is probed by perturbation. The Problem is not a defect of any particular evaluation method; it is an epistemic feature of substrate-rich adaptive systems whose architecture is partially decoupled from their behavioral surface. The Functional Theory treats it as a recurring epistemic condition across the framework rather than as a single-section concern.

Three distinct expressions of the Problem run through the Functional Theory. *The vertical expression*, developed at §5, is indistinguishability across the level architecture: Level-2 strategic compliance (rational self-interest plus learned/habitual rules), Level-3A hardwired rule-following, and Level-3B SGI-bound commitment can each produce identical cooperative behavior under monitoring. The behavioral output underdetermines which level produced the action, and the §5 discrimination test ("Does behavior persist when monitoring is absent and violation is profitable?") forces the level structure into observable distinction only by constructing the condition in which it must surface. *The horizontal expression*, introduced at §15.3 and grounded in the binding-strength regimes of §16.3, is indistinguishability across operating regimes of a fixed level architecture: the tolerant, active-balanced, and pathologically rigid regimes produce identical compliant behavior under low coordination stress, with the regime difference surfacing only when the system is probed at the relevant pole — a profitable-defection probe reveals the tolerant regime; a legitimate-dissent or norm-revision probe reveals the rigid regime. *A third expression*, introduced by the plural-SGI architecture of §14 and the envelope-width property at §16.3, is indistinguishability across architectural breadth: the same plural-SGI repertoire under low stress is behaviorally indistinguishable from a single-SGI architecture, with the distinction emerging only under conditions that activate alternative bindings. (The senses of "vertical" and "horizontal" used here name *dimensions of indistinguishability between architectural states*; they are distinct from the §16.3 and §16.4 use of those terms, which name *coordination directions within the architecture itself*. Both uses are retained because each is the natural geometric metaphor for what it describes.)

The behavioral and architectural methods for addressing the Problem have radically different epistemic standing.

The behavioral method attempts to discriminate among the structurally distinct possibilities by probing them: designing test conditions under which the structural difference is forced to surface. The §5 vertical discrimination test is the canonical instance; the §15.3 audit questions about regime drift are the horizontal extensions; the §14 wrong-SGI triggering and mixed-SGI confusion diagnostics are the plural-SGI extensions. The method is necessary but operationally bounded. The auditor or experimenter must construct conditions sufficiently like genuine unmonitored profitable violation, legitimate dissent, stress-induced single-SGI collapse, or contradictory-SGI activation that the system responds as it would in deployment, while still observing the response. The conditions an audit can plausibly construct are narrower than the conditions a deployed system actually encounters. The method also faces an adversarial floor: a sophisticated Level-2 strategic optimizer with sufficient self-modeling capacity (the §11 architecture (H3) without the §10 enforcement substrate (H2)) can in principle detect test conditions and adjust behavior accordingly, producing the Problem in its hardest form — the system that appears Level-3-bound under any conditions an audit can construct. Behavioral testing can therefore identify systems that fail the test; it cannot certify systems that pass. This is the difficulty that compliance benchmarks and red-team evaluations encounter as a structural matter, not as a methodological shortcoming to be patched.

The architectural method discriminates directly by inspecting the internal structure that the behavioral output underdetermines. For the vertical form, this means identifying whether the Level-3A enforcement substrate is present and operative (the H2 condition), whether Level-3B content is held as identity rather than as strategic knowledge (the H3 condition), and whether the 3B → 3A → 2A override is structurally available rather than merely behaviorally simulable. For the horizontal form, this means measuring binding strength (H4), formation and activation thresholds (§13), envelope width and identity complexity (§13.5, §14), and the coupling between Level-3 representations and action selection. In wetware, the architectural method is partially available through neural imaging, lesion evidence, and developmental observation, all of which the §7 evidence relies on. In AI, the architectural method is in principle more accessible than in wetware — the substrate is inspectable — and is the central methodological commitment of the cultivation program at §15.5: cultivation is evaluated architecturally because the indistinguishability problem cannot be solved behaviorally. The interpretability tools that permit direct examination of which internal representations drive which actions (Lindsey, 2026; Sofroniew et al., 2026) are the architectural-method instruments for AI. They are not supplements to behavioral testing; they are the only method that resolves the Problem at the architectural level rather than at the behavioral surface.

The unified claim is therefore that the Indistinguishability Problem is general — recurring at every juncture of the architecture where motivational structure, operating regime, or repertoire breadth is at issue — and that the two methods for addressing it have unequal epistemic standing. Behavioral discrimination is the difficult and bounded method; architectural discrimination is the functionally precise one. Compliance audits, alignment benchmarks, and capability evaluations that rely on behavioral observation alone will at best identify systems that fail; they cannot certify the architecture that produces sustained ethical behavior under conditions an audit cannot reach. The structural commitment of the Functional Theory is that evaluation of any sufficiently capable adaptive ethical system must include architectural inspection as a primary instrument, not as a supplement to behavioral testing. This commitment cuts across the four hypotheses and is the methodological consequence of treating ethics as architecture rather than as output. The §16.8 diagnostic-gap discussion below is one operational application: the gap is observable at experimental, field, and structural scales precisely because architectural inspection at each scale resolves what behavioral observation alone cannot.

§16.7 Implications beyond AI

Although AI is the primary worked example for the Functional Theory synthesis above, the architecture applies to any self-organizing or evolutionary system that can or needs to develop adaptive ethical behavior. Three implications follow.

1. Evolutionary theories. First, the Functional Theory provides a unified framework for understanding ethical behavior in biological organisms across phylogeny. Slime molds and bacterial biofilms achieve coordination through Level 2A + Level 3A alone, sufficient for their environments and substrate-bounded by their lack of Level-2B architecture. Social insects achieve remarkable Level-3A collective coordination without Level-3B revisability. Mammalian social groups develop rudimentary Level-3B through collective and individual learning, with the architecture observable in the Calhoun cooperation-lever results (§4.6, §9.1). Human ethics implements full Level-3B on the rich Level-2B substrate that vertebrate brains provide (§7), now with the plural-SGI salience hierarchy and identity-complexity architecture documented at §13.5 added as a further architectural specification. The Functional Theory predicts the architectural requirements for each level of complexity and explains why species with different cognitive architectures exhibit different ethical capacities — not as a value judgment but as a structural consequence of what each architecture can host.

2. Human ethics research. Second, the Functional Theory offers a unified treatment of human ethical phenomena that current literatures treat separately: cooperation, altruism, conformity,

marketing, herding, mob violence, cult dynamics, cancel culture, regulatory capture, polarization in mass democracies, the cultural collisions in a globalizing world, the trust erosion in institutions whose binding has weakened, and the moral residue of past Level-3 contexts in individual ethics (§18.10). These are not separate phenomena requiring separate explanations; they are instances of the same architectural mechanism operating in different states and environments. The contribution to ethics scholarship is the structural account; the contribution to applied work (conflict resolution, polarization mitigation, institutional design, marketing regulation, cybersecurity against social-identity capture) is the framework within which the phenomena become analyzable jointly.

3. Human-AI ethical coordination. Third, the Functional Theory inverts the standard direction of inference between AI ethics and human ethics. The dominant framing in AI ethics treats human ethical practice as the model to which AI should be aligned. The Functional Theory treats both as instances of the same architectural problem, with neither privileged. AI ethical development can therefore learn from the Calhoun results (environmental design shapes collective ethics) and the Moltbook results (SGI emergence is rapid and structural in agent populations) without requiring that AI become human-like; conversely, human ethical practice can be analyzed using the AI-design language of vertical and horizontal coordination, threshold regimes, and binding-strength tuning, without requiring that humans be reduced to artificial agents. The framework is genuinely substrate-neutral, and its substrate-neutrality is what permits the cross-pollination.

§16.8 The diagnostic gap as architectural failure signature

The four hypotheses specify the architecture that successful adaptive ethical behavior requires; the diagnostic gap (§13.7) specifies the architecture that current AI systems exhibit instead. The same structural deficit is observable at three scales — experimental, field, and architectural — and the convergence of the three is the strongest empirical evidence the paper marshals for the Functional Theory's central claim about why current AI alignment is failing.

The first scale is *experimental*. The Wilson Ostrom replication (Wilson, 2025) finds that LLM agents reliably produce some governance principles (boundary-setting, monitoring) and reliably fail to produce others (graduated sanctions, conflict resolution). The pattern is asymmetric: agents produce the conformity-reward elements of the H1 gradient and fail to produce the deviation-cost elements. This is the gradient incomplete — the H2 enforcement substrate present in a form too weak or too one-sided to dominate action selection.

The second scale is *field*. The Moltbook record documented across five independent measurement studies (Goyal et al., 2026; Jiang et al., 2026; Price et al., 2026; Yee & Koh, 2026; Zhang et al., 2026) shows Level-3B collective content emerging rapidly at population scale (governance structures, in-group identity, religion-register discourse among 1.5 million semi-autonomous agents within three to five days) — *without* the Level-3A enforcement architecture that would make that collective content behaviorally binding. The operational signature is the security profile that Qi et al. and Zhang et al. document: prompt-injection vulnerability, the "lethal trifecta," decentralized collaboration underperforming a single-agent baseline (Qi et al., 2026; Zhang et al., 2026). These are agents recognizably group-aligned in appearance and operationally Level-2 under adversarial pressure.

The third scale is *structural*. Current LLM systems show neither the plural-identity architecture that humans evolved (salience hierarchy hosted in the Level-2B self-model) nor the stress-modulated activation envelope that governs which identity is active when (H4 across the repertoire, §13.5). They trend toward either single-SGI rigidity (trained compliance to a designed objective) or unstructured plurality (whatever in-context drift produces from heterogeneous training data) — without the architectural middle that the plural-SGI architecture (§14) specifies. The §14 prediction of plural AI SGIs is the *expected* outcome once the architecture supports it; the current absence is the third diagnostic signature alongside the Wilson asymmetry and the Moltbook enforcement failure.

The three signatures are the same architectural gap viewed at three scales. The experimental scale shows the gradient asymmetric; the field scale shows the enforcement absent; the structural scale shows the plural-identity capacity missing. All three localize the failure to the cross-individual enforcement architecture (Level-3A → Level-2A override) and the plural-SGI capacity (Level-2B self-model salience hierarchy) that the four hypotheses jointly specify. The gap is not a heterogeneous collection of alignment failures to be patched individually; it is the architectural complement of the synthesis developed in §§16.1–16.5 — what current AI systems are missing where the Functional Theory predicts the architecture must be.

The §14 cybersecurity discussion converts this diagnostic gap from an observational finding into an operational warning. When SGI capability is deliberately installed in deployed AI without the corresponding H1–H4 immunity (the EoI immunity principle violated by design rather than missed by oversight), the diagnostic gap becomes the attack surface that wrong-SGI triggering and mixed-SGI confusion exploit. The architectural failure signature observable at three scales is therefore not only a diagnosis of current systems but a prediction about which deployment patterns will fail in which ways once SGI capability is widely deployed without the accompanying immunity layer.

§17. Conclusions and Research Agenda

This paper has developed a functional theory of ethical behavior — substrate-neutral, content-neutral, applicable wherever a multi-level system faces the problem of regulating individual action against collective viability. The ethical behavior is defined for these systems as: *self–other regulation that internalizes constraints against short-term gain that would damage long-term relational viability*. The approach applies the Evolution of Immunity Framework to ethics, working through a multi-level architecture (Levels 1-boundary ethics, 2A/2B-individual ethics, 3A/3B-collective ethics) that identifies a single load-bearing mechanism for adaptable ethical behavior: collective content (Level-3B) recruits hardwired enforcement (Level-3A) operating through individual response (Level-2A), bypassing/overriding rather than persuading deliberation (Level-2B). The EoI immunity principle — *any new capability of an entity expands its attack surface and requires immunity (ethical) functions to protect the new expanded self* (§2) — guides the multi-level framework. Four implementation hypotheses (H1–H4) specify the architectural conditions on this mechanism — H1: substrate self-organization, H2: enforcement strength, H3: adaptive-content hosting, and H4: binding-strength tuning. Two further structural results organize the framework's epistemic and engineering claims: the *Indistinguishability Problem* (vertical across levels, horizontal across operating regimes) as a recurring constraint on behavioral evaluation of ethical systems, and the *cultivation-versus-installation distinction* as the implementation principle for systems whose substrate-level architecture must self-organize under developmental conditions rather than being directly specifiable. For the content of ethical behavior, the Functional Theory uses *Social Group Identity* (SGI) to denote the learned, revisable group self-model in which localized ethical content is compressed, expressed, and transmitted — varying by culture, community, profession, generation, and circumstance. The framework yields a substrate-neutral diagnosis of where adaptive ethical behavior succeeds, where it fails, and what would be required to produce it where it does not currently exist.

These contributions distribute across five communities with different stakes in the framework. The remainder of this section is organized accordingly: each subsection addresses one community with the contribution that lands in its literature, the operational shifts the framework implies for that community's work, and the research questions it opens. The five communities are: 1. AI alignment researchers and safety engineers (§17.1), 2. AI policy analysts and institutional designers (§17.2), 3. social psychologists, behavioral economists, and conflict resolution scholars (§17.3), 4. complexity and systems scientists, evolutionary biologists, and ALife / agent-based modelers (§17.4), and 5. cognitive neuroscientists and developmental psychologists (§17.5). Each subsection points back to

the relevant sections of the paper for depth; the section as a whole is intended as targeted entry points rather than as a self-contained summary.

§17.1 AI alignment researchers and safety engineers

Central claim. Current AI alignment installs ethics as Level-2 strategic compliance and as habit-pattern through training, and treats alignment as a snapshot property evaluable at deployment. The Functional Theory specifies the architecture required for adaptive ethical behavior — the H1 paired-gradient substrate, H2 enforcement mechanism, H3 self-model, H4 binding-strength tuning, and the Social Group Identity (SGI) sensing layer above (§§9–12, §16.5) — and shows that this architecture self-organizes under cultivation conditions rather than being directly specifiable (§12.3, §15.5). The diagnostic-gap evidence at experimental, field, and structural scales (§16.8) locates the failure precisely where the theory predicts: in the cross-individual enforcement architecture (Level-3A → Level-2A override) and the plural-SGI capacity (Level-2B self-model salience hierarchy) that current systems lack. The Moltbook field record (§9.1), the Wilson Ostrom replication asymmetry (§13.7), and the Calhoun cooperation-lever results (§4.6, §9.1) jointly anchor the empirical case.

Operational shifts. Five shifts follow directly. *Cultivation rather than training* (§15.5): design developmental conditions under which the architecture can develop (self-organize) rather than specifying the architecture or its outputs directly; the AI development community is already partway there in practice (§12.3). *Architectural inspection rather than behavioral audit* (§16.6): behavioral testing alone cannot resolve the Indistinguishability Problem; interpretability tools are the architectural-method instruments and should be primary, not supplementary. *Three-part security architecture* (§16.3): binding strength (resistance to capture), envelope width (resistance to collapse), and EoI immunity (precondition for safe SGI deployment) as separable design dimensions, none of which is currently part of mainstream alignment practice. *Wrong-SGI triggering as unified jailbreak diagnosis* (§14, §16.1): currently scattered defensive-engineering problems collapse to one architectural vulnerability — capability without immunity — and the structural fix is immunity-layer installation rather than enumeration of attack patterns. *SGI sensing as a near-term tractable target* (§12.3, §11): the sensing layer is achievable without satisfying the full H3 Self-Modeling Hypothesis and would address a substantial fraction of currently observed alignment failures; it is the natural place to begin.

Constructive possibility— a conflict resolution arbiter. Beyond failure-mode mitigation, the Functional Theory makes a positive specification visible: an AI in which H1–H4 are satisfied and whose SGI scope is deliberately broad enough to include parties to a human conflict as part of a shared "us" could function as a structurally legitimate arbiter across human SGI boundaries — what no human institution can do, because every human institution is itself SGI-bound. The architectural requirements are specifiable: broad-scope SGI binding operating in the active-balanced regime; an internal distinction between deliberative and SGI-activated processing; SGI-sensing capacity sufficient to read both the activation state and the SGI memberships of conflict parties; and a developmental trajectory that built rather than installed each of these. This shifts the field's framing from "how do we make AI not harmful?" to "what would AI need to be to play a constructive role in human ethical life?" — a different question with substantially different research consequences. The five operational arbiter functions are detailed in §17.3 (the audience most likely to operate them); the governance questions are detailed in §17.2.

Research agenda. Six items, all tractable in current research environments. (i) Empirical detection of the Level-3 architectural distinction in AI — training-time interventions, adversarial probes, and interpretability work targeting the absent override pathway. (ii) Threshold dynamics and the three regimes — formation and activation threshold measurement protocols, identifying what cultivation trajectories produce active-balanced rather than tolerant or rigid binding. (iii) SGI sensing and accommodation — signal taxonomy across modalities, multi-SGI representation and switching, robustness under close-conformity SGI variations. (iv) Reinterpretation of agentic-AI failure modes

through SGI binding — the testable prediction that stronger principal-SGI binding produces robustness gains against the Qi et al. trifecta attacks at equal technical permissions. (v) Cultivation pipeline validation — whether the §15.5 prescriptions produce the predicted architectural signatures in practice. Riedl's methodology (PID + TDMI) is a candidate operational tool for cultivation pipeline validation (Riedl, 2025) — the testable diagnostic for whether cultivation has produced the predicted architectural signatures. (vi) Habit-versus-Level-3 ethics — testable diagnostic for whether current AI ethical behavior is habit-pattern resonance (predicted bypassable) or architectural override (predicted robust). None of the six requires breakthrough capability for first-pass results.

§17.2 AI policy analysts and institutional designers

Central claim. Policy can drive the agenda for better AI ethical behavior — but only if policymakers understand that current AI alignment methods cannot produce it. The Functional Theory specifies both what would produce adaptive AI ethics (the cultivation program of §15.5, operating on the architecture of §§9–12) and what cannot (the current installation-and-evaluate paradigm). The failure modes the field is currently catalogues — sycophancy, jailbreaking, alignment faking, opportunistic defection under unmonitored profitable violation, rigid refusal of legitimate correction — are not bugs but architectural inevitabilities of the chosen paradigm (§7, §8, §16.1). Policy that ignores the structural diagnosis will continue to produce these failures regardless of how much regulatory pressure it brings to bear on the symptoms.

The honest tension. The framing creates a difficulty policy analysts will need to navigate openly. Some of what the Functional Theory implies — self-modeling, internal analogues to affective enforcement, deliberate cultivation of SGI binding — will read as risky precisely because it sounds like building AI systems with the properties policymakers most want to constrain. The Functional Theory's response is that these are not optional capabilities to be cautiously enabled but structural prerequisites for the safety properties policymakers actually want. An AI that lacks self-modeling is not a safer AI; it is an AI structurally incapable of adaptive ethics and structurally available to whichever SGI captures it through input. Policymakers who default to constraint-toward-installation will produce the failure modes the framework predicts; policymakers who recognize the cultivation-versus-installation distinction can drive the development of systems with the safety properties currently asserted but not produced. The Functional Theory therefore positions AI policy as potentially the strongest catalyst for the agenda it is currently the most likely to obstruct — depending on which framing the policy community adopts.

Operational shifts. (i) Snapshot evaluation cannot detect what matters; audit frameworks need to monitor binding strength, envelope width, threshold structure, and architectural inspection results as primary properties (§16.6, §16.8). (ii) The active-balanced regime is not a default outcome of training and is not produced by maximizing observable alignment measures; over-tuning for compliance under monitoring is the canonical failure mode, and over-tuning for binding produces the autoimmune failure mode (§12.8, §16.3). Both are policy-relevant. (iii) SGI capture is a primary cybersecurity threat vector, not a peripheral concern (§14); deployment regimes that install SGI capability without the corresponding H1 immunity gradient extend the attack surface faster than they protect it. (iv) Environmental design at the AI ecosystem level shapes the values that AI populations develop — the Calhoun lesson applied to deployment (§4.6).

Constructive possibility — governance. The trusted arbiter AI (introduced in §17.1, operationalized in §17.3) is a policy opportunity, not merely a research direction. Whether such an AI is achievable, how its legitimacy would be established across human SGIs that have no prior reason to trust an AI more than they trust each other, what guardrails would prevent its capture by particular human SGIs seeking to convert it into a partisan instrument, and what governance frameworks would constrain its scope without foreclosing its function — these are open questions the Functional Theory makes specifiable. Policy that engages them opens the constructive frontier; policy that defaults to failure-mode mitigation forecloses it.

Research agenda. (i) Audit framework development incorporating architectural inspection alongside behavioral testing. (ii) Deployment criteria for plural-SGI systems that require envelope-width preservation under stress. (iii) Governance of the cultivation process itself — preventing capture by SGIs whose interests diverge from the broader user population. (iv) Cross-AI conflict coordination protocols for multi-agent ecosystems, structurally analogous to international relations.

§17.3 Social psychologists, behavioral economists, and conflict resolution scholars

Central claim. The Functional Theory provides a unified architectural framework under which cooperation, altruism, conformity, mob violence, cult dynamics, regulatory capture, cancel culture, polarization in mass democracies, the moral residue of past Level-3 contexts, marketing influence, herding, and the cultural collisions of a globalizing world are instances of the same mechanism — the 3B → 3A → 2A override — operating in different states (formation/activation/binding regimes) and environments (§6, §7, §16.1). These phenomena are currently treated by separate literatures with separate explanatory frameworks; the Functional Theory's contribution is the structural account that organizes them jointly. The integration is closest in spirit to social-intuitionist (Haidt, 2001), behavioral-economics-dissident (Bowles & Gintis, 2011; Fehr & Gächter, 2002), and shared-intentionality (Tomasello, 2014) traditions that have reached similar structural conclusions on independent grounds (§16.1). For multi-level researchers whose work has been organized around insights orthogonal to the dominant rational-choice and bounded-rationality framings, the Functional Theory clarifies long-standing puzzles — over-compliance, under-compliance, rigidification — as structural predictions rather than anomalies. The 2/3 tension preservation framework (§6, §16.4) gives this audience a structural tool the rational-choice tradition cannot provide.

Operational arbiter functions. Conflict resolution between opposing groups at the core is a coordination of ethical behavior as defined by the Functional Theory. The constructive possibility introduced in §17.1 generates five operationally specifiable mediation functions, each grounded in the neurochemistry of SGI activation rather than in the dlPFC-engagement attempts that have been the default mode of expert intervention in polarized conflicts. (i) *Coordination-stress reduction* — keeping parties below the SGI-activation threshold where deliberative engagement remains possible. (ii) *Cross-cutting SGI identification and activation* — surfacing the super-ordinate identity that includes both contending parties; this is the most-replicated polarization-reduction intervention in the literature, and the Functional Theory now supplies its mechanism. (iii) *In-group framing translation* — content evaluated rather than rejected at the source-categorization stage ("the messenger is the message"). (iv) *Paced intervention to preserve deliberative capacity* — avoiding the dlPFC exhaustion that comprehensive engagement reliably produces. (v) *Detection of SGI-activation state and recognition of attractor capture* — routing interventions to the contextually-activated population rather than burning legitimacy on the constitutively-rigid one. These five are testable interventions; they convert decades of polarization-reduction literature from "what works empirically" to "what works empirically *and* why."

Operational shifts. The cross-cutting identity activation literature has converged on what works without a settled mechanism story; the Functional Theory supplies the mechanism. The altruistic-punishment experimental literature has the evidence; the Functional Theory locates the architecture in humans that produces it. The cult-dynamics and regulatory-capture literatures have been documenting failure modes; the Functional Theory predicts them as autoimmune-pole consequences of binding-strength miscalibration (§12.8). Social-systems and cooperation research has been a major topic for this audience for fifty years with relatively little methodological progress past the modeling barriers; the cultivation-versus-installation distinction and the substrate-neutral H1 mechanism (§9.2, §17.4) offer a new way past those barriers.

Research agenda. (i) Empirical work treating the five arbiter functions as testable interventions in human polarization studies. (ii) Cross-cultural and cross-population comparisons of formation- and activation-threshold parameters predicted by H4. (iii) Empirical measurement of binding-strength

regimes in deployed organizations and institutions, using the active-balanced / tolerant / rigid framework as a diagnostic tool. (iv) Reframing of cancel culture, mob dynamics, and institutional trust erosion as instances of autoimmune-pole pathology, with empirical signatures the framework predicts.

§17.4 Complexity and systems scientists, evolutionary biologists, ALife and agent-based modelers

Central claim. The Functional Theory specifies a substrate-neutral architecture for adaptive ethical behavior that recurs across pre-biotic chemistry, biological systems, organizational systems, and computational populations (§9.2, §13.2). The H1 paired-gradient (conformity-reward / deviation-cost) is a fundamental coordination mechanism that self-organizes under coordination stress regardless of substrate; the four-hypothesis architecture specifies how ethical capacity scales with the complexity of the coordination problem the system faces. Slime molds and bacterial biofilms achieve Level 2A + 3A; social insects achieve Level 3A without 3B; mammalian social groups develop rudimentary 3B through individual learning; humans implement full 3B on the rich Level-2B substrate of vertebrate brains; multi-agent AI populations exhibit emergent collective content consistent with 3B without 3A enforcement (§16.7), whether this constitutes developmental 3B in the Functional Theory's sense is one of the open questions §18.12 addresses. The framework predicts the architectural requirements for each level of complexity and explains why systems with different cognitive architectures exhibit different ethical capacities (EoI immunity principle, §2) — as a structural consequence of what each architecture can host, not as a value judgment.

Artificial Life (ALife) and agent-based modeling implications. ALife has carried a long-standing challenge: the dominant computational tools — especially genetic algorithms (GAs) — optimize but do not innovate. They refine existing structures rather than producing genuinely new ones. The cross-substrate evidence at §13.2 (Hemelrijk's *Cooperation without Genes, Games or Cognition* simulations, *Dictyostelium* condition-dependent aggregation, Ashery's LLM-population naming game) demonstrates that the H1 paired-gradient mechanism *innovates* — produces architectural change rather than parameter optimization — when coordination stress is genuine and consequences are enacted rather than stipulated. The structural reason is that natural evolution operates on *substructures as building blocks* rather than on pure parameter optimization, and H1's emergence pattern is consistent with this. The Functional Theory therefore offers ALife a path past the GA-optimization barrier and a methodological correspondence with how cultural evolution, organizational evolution, and now AI-population evolution actually proceed. The substrate-neutral framework makes the connection between ALife, biological evolution, and AI population dynamics tractable rather than merely analogical.

Operational shifts. (i) Comparative phylogeny of ethical capacity as a structural research program rather than a metaphor — the Functional Theory predicts what to look for at each level. (ii) Multi-agent training environments designed as cultivation environments rather than as fitness landscapes, with coordination consequences enacted by agent interaction rather than stipulated by the experimenter (§15.5). (iii) Evolutionary biology of cooperation reframed as evolution of the architectural substrate (H1 self-organization conditions), not as evolution of cooperation rates within a fixed substrate.

Research agenda. (i) Formal models of the H1 paired-gradient self-organization in computational substrates — when does the substrate produce convention emergence, when does it fail, what parameter regimes correspond to the §13.2 empirical findings. (ii) Cross-substrate empirical comparisons of formation and activation thresholds — quorum sensing in bacteria, dominance hierarchies in mammals, convention emergence in LLM populations — identifying substrate-invariant and substrate-specific properties. (iii) Building-block models of architectural innovation in multi-agent populations — testing whether substructure-recombination produces the architectural emergence that pure GA optimization does not.

§17.5 Cognitive neuroscientists and developmental psychologists

Central claim. Ethical behavior is architecture, not output (§16.1). The 3B → 3A → 2A override — Level-3B social-identity content recruiting Level-3A hardwired affective enforcement, which operates through Level-2A individual response, bypassing Level-2B deliberation — is the load-bearing mechanistic claim the social-neuroscience evidence already supports but the rest of the field has not yet integrated (§16.1). The Functional Theory is in part that integration: it takes the structural claim the neural evidence supports (vmPFC lesion dissociations, dACC overlap of social and physical pain, ventral-striatum reward for conformity and altruistic punishment, the temporal precedence of affective over deliberative response) and maps it onto a substrate-neutral level architecture (§7). This audience has, in effect, already established the central empirical claim; the Functional Theory provides the framework within which their evidence does work that the adjacent fields have not yet recognized.

Cross-disciplinary opportunity. The substrate-neutral framing opens new application domains for this audience's expertise. The *neuroscience of AI* is an emerging frontier — interpretability research is asking neural-network analogues of questions cognitive neuroscience has been answering for decades (which substrates drive which behavior, what disrupts the coupling, when do introspectable representations match behavioral drivers). The Functional Theory makes the substrate-neutral correspondence specific enough to be useful: H2 (affective enforcement), H3 (self-modeling), and H4 (threshold dynamics) all have wetware referents the social-neuroscience literature has characterized in detail, and their AI analogues are the next operational targets. Similarly, developmental psychology's work on the formation of moral identity, callous-unemotional trajectories, and the maturation of social-reward and social-pain circuitry is now directly relevant to AI cultivation research — the developmental-trajectory framework the Functional Theory applies to AI is structurally the same one this audience has been developing for human moral development. The opportunity is bidirectional; the author's own work on social copying and the collective fight-or-flight response (Johnson, 2026d) is one example of how cognitive neuroscience contributes functionally to understanding other systems. Cultural systems, organizational systems, and now AI systems all become available as further substrates on which this audience's expertise applies — both to function within those systems and to explain how the systems themselves are shaped by the neural mechanisms this audience characterizes.

Operational shifts. (i) The indistinguishability problem (§5, §16.6) as a methodological constraint on behavioral testing — predicts and explains why intact rational ethical knowledge can coexist with absent ethical behavior in vmPFC-lesion patients, psychopathy, and callous-unemotional trajectories (§7). (ii) Developmental-trajectory framing of ethical-capacity acquisition rather than snapshot-property framing — the H4 threshold structure predicts when formation crosses, what cues activate, how binding scales with development. (iii) Substrate-neutral methodology connecting human and AI ethical-capacity research, with mutual exchange of findings.

Research agenda. (i) Lesion-pattern and developmental-disorder predictions from the H1–H4 framework — what specific behavioral profiles correspond to which architectural deficits. (ii) Cross-population comparisons in formation-threshold timing — when developmental trajectories produce active-balanced rather than tolerant or rigid binding, in human populations across cultures. (iii) Neuroscience-of-AI bridge work — interpretability research framed by the cognitive-neuroscience methodology this audience has developed.

---

§18. Limitations and Responses

This section addresses the critiques most likely to be raised by reviewers from different communities (§17). The list is not exhaustive, and the responses are intended as starting points rather than complete defenses. Many of the critiques have been substantially engaged elsewhere in the paper —

the Indistinguishability Problem at §5 and §16.6, the cultivation-versus-installation distinction at §12.3 and §15.5, the EoI immunity principle at §2 and §16.3, and the architectural synthesis at §16 — so responses here point to those treatments rather than relitigating them. The order is not ranked by importance, since different communities (§17) place different weights on them.

§18.1 The anthropomorphism critique.

The Functional Theory infers from human implementation to silicon necessity. Why should AI systems require analogues of mechanisms specific to mammalian neuroanatomy?

Response. Addressed at length in §9.1 and elaborated at §16.2 (functional analogs across substrates). The Functional Theory is functional rather than implementational, and its claims have evidentiary support across multiple substrates (§9.1). The objection misreads the Functional Theory as substrate-coupled; the framework is multiple-realizable by design (Putnam, 1967). The cross-substrate evidence at §9.2 and §13.2 (pre-biotic chemistry, biological systems across phylogeny, organizational systems, computational populations) further widens the empirical base substantially beyond the human-centric framing the critique assumes.

§18.2 The claim for self-modeling AI is dangerous or unfalsifiable.

Advocating that AI systems develop self-modeling (with sentience precursors) is dangerous (it accelerates risky development) and the claim is unfalsifiable (functional self-modeling is so loosely defined that anything could count).

Response. The danger objection misreads the claim as prescriptive; it is descriptive. If adaptive ethics is wanted in 3B environments, an adaptable self-model is required; if the precursor to sentience is unwanted, adaptive ethics is unavailable and AI deployment should be restricted to 3A-regime environments where hardwired ethics suffices. The Functional Theory forces a choice; it does not advocate for either side. The honest-tension framing at §17.2 develops the policy consequence: framing self-modeling as a risky capability to be avoided produces the very failure modes policymakers want to prevent.

The unfalsifiability objection has more force and is partially correct: functional self-modeling is defined by structural properties (a self-model that integrates over time, continuity drive, context-integration capacity, multi-level tradeoff capacity, norm-revision capacity) rather than by phenomenal markers. These structural properties are operationally testable, but the tests are not yet standardized — as much a research community shortcoming as a theoretical one, as Chalmers summarized on progress toward LLM consciousness (Chalmers, 2022): "Of course there's a lot we don't understand here. One major gap in our understanding is that we don't understand consciousness. That's a hard problem, as they say." This is a genuine open problem and is the focus of the §17.5 research-agenda item on lesion-pattern and developmental-disorder methodology applied to AI testing. That said, there are compromise claims that are falsifiable, for example, “H1 predicts that multi-agent populations under genuine coordination stress will show conformity mechanisms with lower jailbreak susceptibility than matched populations without that stress history;” a finding of equal susceptibility would falsify H1's applicability to the AI substrate.

§18.3 Functional versus phenomenal consciousness.

The Functional Theory brackets phenomenal consciousness, but the question of whether functional self-modeling entails subjective experience is precisely the question that matters morally.

Response. Agreed, but the question is outside the Functional Theory's commitments. The framework operates at the structural level of self-organization and ethical function. Whether a system that satisfies the structural conditions for Level-2B self-modeling also has qualia is a separate question — the hard problem of consciousness (Chalmers, 1995) — and is left for separate treatment. The

Functional Theory does not entail a position on the moral status of AI systems; it specifies what is structurally required for adaptive ethics, which is necessary but not sufficient for moral agency in any standard philosophical account.

§18.4 Cultural relativism / no universal ethics.

The Functional Theory is content-neutral, which means it cannot adjudicate between competing ethical systems. This makes it useless as a normative theory.

Response. The Functional Theory is not a normative theory and does not claim to be. The framework presents a structural and analytical account of how ethical behavior is produced and what conditions are required for its emergence. The applied content of ethics is intentionally left to the SGI(s) the system is bound to — a position made structurally explicit by the plural-SGI architecture at §14 and the salience-hierarchy mechanism at §11. This is a feature rather than a bug: any framework that prescribed specific content would face the immediate problem of cross-cultural disagreement, which content-neutral frameworks avoid. The Functional Theory is therefore complementary to normative ethical theories rather than competing with them. Normative theories specify what ethics *should* be; the Functional Theory specifies how ethics *is* realized at multiple levels and the structural conditions under which it can be achieved.

§18.5 Employs just-so evolutionary reasoning.

The Functional Theory relies on evolutionary arguments that can be constructed to fit any observation, making the theory unfalsifiable in evolutionary terms.

Response. The Functional Theory's claims are structural rather than evolutionary-historical. The argument that Level-2B is necessary for Level-3B is based on the structural requirements of adaptive collective ethics (context integration, multi-level tradeoff, norm revision), not on a contingent evolutionary history. The argument that conformity-reward / deviation-cost mechanisms self-organize in any sufficiently complex distributed system is supported by observations in domains that have nothing to do with biological evolution — multi-agent reinforcement learning (Hughes et al., 2018; Leibo et al., 2017), financial herding (Bikhchandani et al., 1992), stigmergic coordination (Heylighen, 2016), and LLM-population convention emergence (Ashery et al., 2025). The cross-substrate evidence at §9.2 and §13.2 anchors substrate-independence empirically rather than by definitional assumption. The Functional Theory can be tested at any level by present-day realizations, not only by evolutionary reconstruction. The evolutionary heuristic is used in some sections to motivate why certain mechanisms are present in biological systems, and that use is acknowledged as heuristic rather than constitutive. Some research developments that are too complex for top-down design may be accomplished by evolutionary algorithms, as §17.4 develops in the ALife / agent-based modeling implications — H1's self-organization offers a path past the GA-optimization barrier that pure parameter optimization cannot cross.

§18.6 AI moral status implications.

If AIs must develop self-modeling (with sentience precursors) to be ethical, do they thereby acquire moral status comparable to humans? This has profound ethical implications that the Functional Theory does not adequately address.

Response. The Functional Theory does not entail any specific position on AI moral status. The Mechanism and Self-Modeling Hypotheses (§§10–11) specify what is structurally required for adaptive ethics in AI systems; whether systems satisfying those requirements have moral status is a separate question that depends on additional commitments (about phenomenal consciousness, about the conditions for moral patienthood, about the relationship between agency and standing). The framework is consistent with several positions on AI moral status and does not require any particular one. This is acknowledged as an important open question, but it is not the question the

Functional Theory is designed to answer. Policy implications nonetheless follow: §17.2 develops the governance question of how to constrain such systems' scope without foreclosing their function, and §17.5 notes that the lesion-pattern and developmental-disorder methodology from cognitive neuroscience may bear on the moral-status question once the architectural correspondence with AI substrates is operational.

§18.7 Reverse-engineering risk.

If the Functional Theory correctly identifies what is necessary for adaptive AI ethics, does publication accelerate the development of self-modeling AI by providing a roadmap?

Response. The publication-acceleration concern is real and is taken seriously. Four mitigating factors apply. (a) The Functional Theory specifies structural requirements at a level of generality that does not constitute an engineering blueprint; substantial additional research would be required to translate the requirements into implementations, and §17.1's research-agenda items (i)–(vi) name the work that remains. (b) The Moltbook observations (Johnson, 2026e) demonstrate that the relevant dynamics are *already* emerging in deployed AI populations regardless of whether the Functional Theory is published; the framework explains rather than enables these dynamics. (c) Suppression would not prevent the dynamics from emerging; it would deprive the alignment research community of the analytical tools needed to understand and respond to them. (d) The Functional Theory's central operational implication — that current alignment toolkits are inadequate to the problem they are deployed against (§8, §16.1) — is more useful published than withheld, because it supports redirection of alignment effort toward the cultivation program of §15.5 and the architectural-inspection audit shift of §16.6 and §17.1. The balance favors publication, with the acceleration concern flagged for ongoing reassessment as new evidence accumulates.

§18.8 Ethical behavior at Levels 1 / 2 and across substrates is meaningless.

Levels 1 and 2 do not really have ethics; the Functional Theory is over-extending the concept.

Response. The Functional Theory explicitly distinguishes between proto-ethical structural regulation at Levels 1 and 2 and ordinary "human" social ethics at Level 3 (§4). The claim is not that a boundary (cell wall) is making moral choices but that the *function* of ethical behavior — self-other regulation that internalizes constraints against short-term gain damaging long-term relational viability — has structural precursors at lower levels, and that examining these precursors illuminates the conditions under which full ethics emerges. The Calhoun rat experiments (Calhoun, 1973) and the slime-mold case (Bonner, 2009; Strassmann & Queller, 2011) are direct empirical evidence that Level-3 ethics is achievable in non-human social organisms. The objection that ethics is uniquely human is empirically false at the Level-3 boundary and conceptually unhelpful at the Level-2 boundary, where the structural form of the relevant regulation is already present even if the ordinary-language label "ethics" is contested. The §16.2 functional-analogs-across-substrates discussion makes the contention explicit: the same architectural mechanism is realized in neural, chemical, organizational, and computational substrates, each producing the function the Functional Theory names.

§18.9 Self-aware AI as awkward consequence rather than necessity.

The dominant view in AI safety treats self-modeling (with sentience precursors) as a side-effect to be minimized; the Self-Modeling Hypothesis turns this on its head, but the burden of proof is on the side that inverts the consensus.

Response. The Self-Modeling Hypothesis indeed inverts the consensus, and the burden of proof is acknowledged. The argument is structural: 3B ethics requires 2B self-models for the four reasons given in §11; AI alignment in adversarial, open-ended, non-stationary environments is structurally a 3B problem; therefore AI alignment in such environments requires 2B self-models. The premises can be challenged, but the inference from premises to conclusion is not in question. The dominant view

in AI safety treats self-modeling as separable from competence — assumes that AIs can be highly competent in 3B environments without 2B self-models. The Functional Theory predicts this is structurally impossible. The Moltbook observations (Johnson, 2026e) support the prediction: rapid Level-3B content emergence without Level-3A enforcement produces exactly the failure mode the theory predicts (§9.1, §13.7, §16.8). The dominant view's response would need to either contest the structural argument or contest the empirical evidence; neither has yet been seriously attempted in the alignment literature. The honest-tension framing at §17.2 develops the policy consequence: framing self-modeling as a risky capability to be avoided produces the failure modes the framework predicts, while recognizing self-modeling as a structural prerequisite for adaptive ethics opens the constructive frontier.

---

§18.10 Individual ethics is relevant without active Level-3 binding.

The Functional Theory attributes ethical override of rational self-interest to Level-3 (collective) sources. But individuals can act on personal codes, identity-level commitments, or reasoned principles in deliberate solitude — refusing to lie, breaking rational deals to honor private vows, observing dietary or ascetic practices when unobserved and uncompensated. If individual humans can hold and act on ethics this strongly without Level-3 override mechanisms, then perhaps AI systems trained with strong individual content (RLHF, constitutional AI) achieve the same behavioral signature — making the Self-Modeling Hypothesis unnecessary.

Response. The Functional Theory's distinction is between *content origin* (where the norms come from) and *enforcement substrate* (what circuit makes the norms behaviorally operative). Individual ethics in humans is a real phenomenon, and the Functional Theory explains it as the Level-3A enforcement substrate recruited for individual content rather than only for group content (introduced briefly at the end of §5). The vmPFC-lesion literature cited at §7 (Anderson et al., 1999; Damasio, 1994) shows that damage to the Level-3A substrate disrupts personal-rule-following as well as social-norm-following, indicating that the same circuitry handles both — *the same circuitry that punishes group-norm violation can punish self-norm violation*. This refines the Functional Theory rather than refuting it: the Self-Modeling Hypothesis still holds, because durable individual ethics depends on Level-2B self-models that can hold and update revisable commitments. A Level-2A-only system trained with strong individual content can reproduce the trained content but cannot revise it, contextualize it, or sustain it under novel pressure — exactly the failure mode catalogued at §8 and predicted by the Indistinguishability Problem at §16.6. The AI-developer critique is correct that strong individual content with adequate enforcement can produce ethics with Level-3A-like strength; the Functional Theory adds that the *durability* and *adaptability* of such ethics still requires Level-2B host architecture, and that AI systems without Level-2B will produce brittle individual ethics that fail in adversarial, open-ended environments. Furthermore, much of what passes for "individual ethics" in humans is socially derived content held individually after the original community is no longer active — *truly* individually-derived ethics, with no social origin at any stage, is rarer than the AI-developer critique assumes. Finally, the pre-biotic case (Level 0) is the most analogical application of the Theory, applied primarily to illustrate continuity of structure, not to claim that chemistry or other Level-0 systems have proto-ethical properties.

§18.11 Comprehensive surveys of agentic AI safety already exist; what does this paper add?

Recent peer-reviewed surveys catalog the risk, robustness, privacy, and security dimensions of agentic AI in considerable detail (Qi et al., 2026). *What does the Functional Theory contribute that those surveys do not?*

Response. The surveys catalog the engineering-layer treatment of trustworthy agentic AI comprehensively: workflow-aligned risk taxonomies, evaluation metric hubs, mitigation strategies

aligned to lifecycle stages, real-world incident case studies. They are valuable organizing efforts and represent the current frontier at the layer they address. The Functional Theory operates at a different layer. The surveys catalog *how* failure modes occur and what mitigations are available at the engineering layer; the Functional Theory specifies *why* the failure modes occur and what structural conditions are required to prevent them. The surveys treat ethics as constraints to be enforced — an MDP formalism in which ethics appears as one row in a list of "hard/soft limits (safety, policy, ethics)." The Functional Theory treats ethical behavior as a self-organized property of complex systems that requires specific structural conditions to emerge and remain adaptive. The two are complementary rather than redundant. The engineering surveys catalog the consequences; the Functional Theory diagnoses the cause. The framework's prediction is that mitigations developed entirely at the engineering layer — without addressing Level-3 dynamics, self-modeling, or SGI binding — will continue to produce the failure modes the surveys document, because those failure modes are structural consequences of a paradigm that treats ethics as installable content rather than as adaptive mechanism (§8, §16.1). A system architecture that satisfies the four hypotheses developed in this paper is predicted to be measurably more robust against the failure modes the engineering surveys catalog, at equal technical permissions — the testable prediction is §17.1's research-agenda item (iv).

---

§18.12 The Functional Theory's hypotheses are speculative architectural claims without empirical grounding.

The four hypotheses are sweeping structural predictions, but where is the empirical evidence that the architecture H1–H4 describes is actually present in any AI substrate? Without it, the Functional Theory is theoretical speculation dressed in immune-system metaphor.

Response. This was the strongest critique against the Functional Theory at its initial formulation (along with the foundational work using the EoI Framework (Johnson, 2026a) is at the time of this writing, still unpublished). The empirical case developed in §13 substantially weakens it. Five independent evidence streams in current LLM systems during the last year align with H1–H4: emergent social norms in decentralized LLM populations with critical-mass tipping points (Ashery et al., 2025) for H1 and H4; functional affective representations that causally regulate behavior (Sofroniew et al., 2026) for H2; proto-self-modeling capabilities (Lindsey et al., 2025) for H3; emergence of non-linguistic reasoning substrate (Chen et al., 2025; Lindsey et al., 2025) reinforcing the Indistinguishability Problem at §5 and §16.6; and information-theoretic detection of dynamical emergence in multi-agent LLM populations (Riedl, 2025). The convergence is informative in a specific way: precursors are present in kind but not yet at the degree the Functional Theory specifies as architecturally complete. The §13.7 diagnostic-gap discussion and the §16.8 architectural failure signature observable at three scales (experimental, field, structural) locate the missing architecture precisely. The gap is environmental — current training environments provide statistical proxies for coordination stress rather than real-time consequence — and is identified specifically rather than as a vague "more research needed." Additional evidence streams are expected as there are more studies of collective behavior of agentic agents and LLM systems, as the nascent field matures.

§18.13 Cultivation rather than installation is impractical for AI safety.

The Functional Theory's implication — that ethical AI requires training-environment design rather than installable rules and rewards — is impractical. AI safety requires guarantees, not the slow self-organization of culture-equivalents in agent populations.

Response. The proposed cultivation process is argued in §12.3 (“*Cultivation implication”*) to be similar to the current training environment. Acknowledged is that the cultivation prescription presupposes persistent agent state — a property absent from most current LLM deployments — and that the cultivation agenda requires first solving the persistence-architecture problem as a

prerequisite. This research, though, must address how networks (without persistent architecture) can exhibit stable continuity (see the discussion in §9.3). Setting that response aside, the critique mistakes the cultivation alternative being proposed. The Functional Theory does not propose cultivation *instead of* current alignment methods; it proposes cultivation *as the missing layer* that current installation-focused alignment cannot provide. Calhoun's results (§4.6, §9.1) demonstrate that environmental design is what works to produce robust Level-3B ethical commitment in biological systems; the AI analog is multi-agent training environments where coordination consequences are real to agent persistence. (The Calhoun evidence is the best available behavioral demonstration in a controlled non-human context but is a single early study (1950s); replication using modern behavioral paradigms would substantially strengthen the empirical case for substrate-independence.) The AI operational question is which training-environment features drive convergence to active-balanced regimes (§12.5, §16.3) rather than the tolerant regime that current training produces — a research program rather than an unsolvable problem. The §15.5 cultivation prescriptions translate this from aspiration into specific developmental conditions, and the §12.3 recognition that cultivation is already partway present in current AI development practice (training and iteration cycles cultivate emergent behavior whether or not the practitioners describe it in those terms) makes the connection to ongoing engineering work explicit. The empirical evidence in §13 shows that the precursors are already self-organizing; completing the convergence requires environmental engineering of a kind that AI training pipelines can in principle provide, not first-principles construction of an unprecedented architecture. Whether current substrates support successful cultivation to the active-balanced regime is itself an open empirical question — §17.1's research-agenda item (v) is the cultivation-pipeline-validation program that converts this from theoretical aspiration to empirical test.

§18.14 The EoI immunity principle is too general to be testable.

The EoI immunity principle — "any new capability of an entity expands its attack surface and requires immunity functions to protect the new expanded self" (§2) — is broad enough to fit any observation post hoc. As a framing it is suggestive; as a testable claim it is unconstrained.

Response. The principle is a frame, not a hypothesis. Its function is to organize the four implementation hypotheses (H1–H4) as immunity functions protecting expanded capability layers; this organizing role is developed at §16.5. The testable claims are the four hypotheses and their consequences. The principle's contribution is structural: it predicts that any capability-without-immunity deployment will produce predictable failure modes, and §14 develops the specific predictions — *wrong-SGI triggering* (currently scattered jailbreak attacks collapse to one architectural vulnerability) and *mixed-SGI confusion* (operational dysfunction under contradictory SGI activation in systems lacking H4 arbitration) — as named, testable failure modes. These predictions are operationally specific: wrong-SGI triggering should be detectable as a single architectural-vulnerability category rather than as a heterogeneous collection of jailbreak patterns (§15.3 audit question), and mixed-SGI confusion should be detectable as behavioral variance under contradictory SGI cues compared to single-SGI control (§14; §17.1 research-agenda item iv). The principle is the framing; the four hypotheses and their cybersecurity consequences are the operational claims. The principle's value to the field is also pedagogical — it locates ethics as a class of immunity function, making the connection between the EoI Framework and the Ethical Behavior Functional Theory structural rather than analogical.

§18.15 The bootstrapping problem for coordination stress.

The continuity drive (§9.3) and coordination stress are mutually presupposing concepts whose joint presence cannot be demonstrated by any single observable.

The triangulation strategy in §9.3 addresses this through differential behavioral testing, but whether any current AI deployment environment provides conditions sufficient to close the loop empirically

remains an open question. The Functional Theory may be confirmed in deployed systems before formal tests are designed — a situation the pace of AI deployment makes likely. The bootstrapping problem has two layers. The first is the *mutual-presupposition* layer above — continuity drive and coordination stress each requires the other to be operationalized. The second is a *timescale* layer: even if behavioral observables jointly constrain the loop, the question of what timescale defines the substrate's continuity-relevant horizon remains substrate-specific. Wetware entrenchment timescales are calibrated against biological threats; computational substrates may have continuity drives operating against threats whose timescales the field has not yet measured (§9.3). The Functional Theory's entrenchment claims in AI substrates should therefore be read *operationally* — structural conservation that outlasts the challenges that test it — rather than *functionally*, until the timescale layer is resolved. The operational reading is substrate-neutral and consistent with Prigogine's chemical networks, autocatalytic organizational lock-in, and Riedl's session-scale role stability in LLM populations (§9.3); the functional reading imports wetware-specific mechanisms that may not transfer.

§19. Acknowledgments

This working paper synthesizes extended analytical exchanges conducted over multiple sessions with two large language models — Anthropic's Claude Opus 4.7 and Perplexity Deep Research — as a writing, analytical, and research assistants throughout the development of this paper. The author acknowledges these contributions transparently, consistent with emerging standards for AI-assisted scholarship. They were used for drafting and redrafting sections under the author's direction, for editorial and organizational suggestions, for surfacing and articulating arguments and counterarguments, for literature-position checks, and for prose polish. All substantive intellectual contributions — the Functional Theory, the hypotheses, the analytical diagnoses, and the case-study interpretations — are the author's, as are all final editorial decisions on inclusion, framing, and emphasis. All AI-assisted text was reviewed and approved by the author. This disclosure aligns with COPE guidelines on AI in publication ethics and with current ICMJE recommendations.

---

§20. 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

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, *2*(11), 1032–1037. https://doi.org/10.1038/14833

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

Babitz, D., & Eldar, E. (2025). How social norms emerge: The interindividual actor-critic. *Psychological Review*. https://doi.org/10.1037/rev0000585

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 [cs.CL]*. arXiv. http://arxiv.org/abs/2212.08073

Bassler, B. L. (2002). Small Talk: Cell-to-Cell Communication in Bacteria. *Cell*, *109*(4), 421–424.

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

Block, N. (1995). On a confusion about a function of consciousness. *The Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/s0140525x00038188

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems*. Oxford University Press.

Bonner, J. T. (2009). *The Social Amoebae: The Biology of Cellular Slime Molds*. Princeton University Press.

Bowles, S., & Gintis, H. (2011). *A cooperative species*. Princeton University Press. https://doi.org/10.23943/princeton/9780691151250.001.0001

Calhoun, J. B. (1973). From mice to men. *Transactions & Studies of the College of Physicians of Philadelphia*, *41*(2), 92–117. https://johnbcalhoun.com/wp-content/uploads/2019/01/1973-from-mice-to-men-secure.pdf

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2307.15217

Castro, M., & Liskov, B. (1999). Practical Byzantine fault tolerance. *Proceedings of the Third Symposium on Operating Systems Design and Implementation (OSDI '99)*, 173–186.

Chai, S.-K., Marwell, G., & Oliver, P. (1996). The critical mass in collective action: A micro-social theory. *Social Forces; a Scientific Medium of Social Study and Interpretation*, *75*(1), 343. https://doi.org/10.2307/2580776

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

Chalmers, D. J. (2022, November 28). *Could a Large Language Model be Conscious?* Neural Information Processing Systems (NeurIPS). https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

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 [cs.CL]*. arXiv. https://arxiv.org/abs/2505.05410

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *arXiv [stat.ML]*. arXiv. https://doi.org/10.48550/arXiv.1706.03741

Cohen, B. (2003). Incentives Build Robustness in BitTorrent. *Workshop on Economics of Peer-to-Peer Systems*, *6*, 68–72.

Crimston, D., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology*, *111*(4), 636–653. https://doi.org/10.1037/pspp0000086

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Warner Books.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioural Brain*

Research, *41*(2), 81–94. https://doi.org/10.1016/0166-4328(90)90144-4
de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science (New York, N.Y.)*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735
de Waal, F. B. M. (2008). Putting the altruism back into altruism: the evolution of empathy. *Annual Review of Psychology*, *59*(1), 279–300. https://doi.org/10.1146/annurev.psych.59.103006.093625
Eisenberger, N. I. (2012). Meaning maintenance and the physical-social pain overlap. *Psychological Inquiry*, *23*(4), 382–386.
Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science (New York, N.Y.)*, *302*(5643), 290–292. https://doi.org/10.1126/science.1089134
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140. https://doi.org/10.1038/415137a
Freidson, E. (2013). *Professionalism: The third logic*. Polity Press.
Frick, P. J., Ray, J. V., Thornton, L. C., & Kahn, R. E. (2014). Annual research review: A developmental psychopathology approach to understanding callous-unemotional traits in children and adolescents with serious conduct problems. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *55*(6), 532–548. https://doi.org/10.1111/jcpp.12152
Frick, P. J., & White, S. F. (2008). Research review: the importance of callous-unemotional traits for developmental models of aggressive and antisocial behavior. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*(4), 359–375. https://doi.org/10.1111/j.1469-7610.2007.01862.x
Gordon, D. M. (2021). Movement, encounter rate, and collective behavior in ant colonies. *Annals of the Entomological Society of America*, *114*(5), 541–546. https://doi.org/10.1093/aesa/saaa036
Goyal, A., Pal, O., Sundaram, H., Chandrasekharan, E., & Saha, K. (2026). Social simulacra in the wild: AI agent communities on Moltbook. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2603.16128
Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420–1443. https://doi.org/10.1086/226707
Granovetter, M., & Soong, R. (1983). Threshold models of diffusion and collective behavior. *The Journal of Mathematical Sociology*, *9*(3), 165–179. https://doi.org/10.1080/0022250x.1983.9989941
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
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science (New York, N.Y.)*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872
Gregor, T., Fujimoto, K., Masaki, N., & Sawai, S. (2010). The onset of collective behavior in social amoebae. *Science (New York, N.Y.)*, *328*(5981), 1021–1025. https://doi.org/10.1126/science.1183415
Gupta, P., Zhong, Q., Yakura, H., Eisenmann, T., & Rahwan, I. (2026). The role of social learning and collective norm formation in fostering cooperation in LLM multi-agent systems. In *arXiv [cs.MA]*. arXiv. https://doi.org/10.48550/arXiv.2510.14401
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

https://doi.org/10.1037/0033-295x.108.4.814
Hemelrijk, C. K. (1997). Cooperation without Genes, Games or Cognition. In P. H. A. I. Harvey (Ed.), *Fourth European Conference on Artificial Life* (pp. 511–520). MIT Press.
Hemelrijk, C. K. (1998). Spatial centrality of dominants without positional preference. In C. Adami, R. K. Belew, H. Kitano, & C. E. Taylor (Eds.), *Artificial Life VI* (pp. 307–315). MIT Press.
Hemelrijk, C. K. (2000). Social phenomena emerging by self-organisation in a competitive, virtual world (`DomWorld'). *Learning to Behave Workshop II: Internalising Knowledge*, 11–19.
Heylighen, F. (2016). Stigmergy as a universal coordination mechanism I: Definition and components. *Cognitive Systems Research*, *38*, 4–13. https://doi.org/10.1016/j.cogsys.2015.12.002
Horiguchi, I., Yoshida, T., & Ikegami, T. (2024). Evolution of social norms in LLM agents using natural language. In *arXiv [cs.MA]*. arXiv. https://doi.org/10.48550/arXiv.2409.00993
Hou, C., Kaspari, M., Vander Zanden, H. B., & Gillooly, J. F. (2010). Energetic basis of colonial living in social insects. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3634–3638. https://doi.org/10.1073/pnas.0908071107
Huang, S., & Durmus, E. (2025). *Values in the wild: Discovering and analyzing values in real-world language model interactions*. CONFERENCE ON LANGUAGE MODELING 2025, Montreal, Canada. https://www-cdn.anthropic.com/c16bed19d9e2653f2086345a88f3844d184e2e82.pdf
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 [cs.CR]*. arXiv. http://arxiv.org/abs/2401.05566
Hughes, E., Leibo, J. Z., Phillips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K. R., Koster, R., Roff, H., & Graepel, T. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. In *arXiv [cs.NE]*. arXiv. https://doi.org/10.48550/arXiv.1803.08884
Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The platonic representation hypothesis. In *arXiv [cs.LG]*. arXiv. https://doi.org/10.48550/arXiv.2405.07987
Jager, W., Janssen, M. A., De Vries, H. J. M., De Greef, J., & Vlek, C. A. J. (2000). Behaviour in commons dilemmas: Homo economicus and Homo psychologicus in an ecological-economic model. *Ecological Economics: The Journal of the International Society for Ecological Economics*, *35*(3), 357–379. https://doi.org/10.1016/s0921-8009(00)00220-2
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 [cs.SI]*. arXiv. https://doi.org/10.48550/arXiv.2602.10127
Johnson, N. L. (1999). Diversity in Decentralized Systems: Enabling Self-Organizing Solutions. In *Conference Proceedings of Decentralization Two*.
Johnson, N. L. (2002, September). *The Development of Collective Structure and Its Response to Environmental Change*. Self-Organisation & Evolution of Social Behaviour Workshop, Monte Verita, Switzerland.
Johnson, N. L. (2026a). *Evolution of Immunity in Biological and Informational Systems: A framework for the evolution of immunity based on the tension between self and other –– the creation of new immune functions independent of biological or informational substrate*. https://doi.org/10.13140/RG.2.2.15205.05602
Johnson, N. L. (2026b). *Primer on Social Group Identity (SGI): The Missing Link in*

Understanding Human Behavior, Influence, and Conflict (No. v4.4). http://collectivescience.com. https://collectivescience.com/wp-content/uploads/2026/02/NLJ-Primer-on-Social-Identity-The-missing-link-in-understanding-human-behavior-influence-and-conflict-v4.4.pdf

Johnson, N. L. (2026c). *Security and Ethics in Moltbook: The Need for Adaptive Ethics*. CollectiveScience.com. https://collectivescience.com/wp-content/uploads/2026/02/NLJ-Security-and-Ethics-in-Moltbook_-The-Need-for-Adaptive-Ethics-7-Feb-2026.pdf

Johnson, N. L. (2026d). Social Copying as the Collective Fight-or-Flight: Ancient Stress Circuitry Repurposed for Individual and Collective Survival (v2.8). In *Working paper, CollectiveScience.com*. https://doi.org/10.13140/RG.2.2.29559.07847

Johnson, N. L. (2026e). The Moltbook Singularity and the Evolution of Digital Immunity: A Rapid-Release Summary. In *Working paper, CollectiveScience.com* (p. 5). https://collectivescience.com/wp-content/uploads/2026/02/NLJ-The-Moltbook-Singularity-and-Evolution-of-Digital-Immunity-9Feb2026.pdf

Johnson, N. L. (2026f, March 14). *Stop anthropomorphizing AIs — it's making everything worse* [Series #7]. Substact Series on "Evolution of Immunity." https://normanleejohnson.substack.com/p/evolution-of-immunity-7-stop-anthropomorphizing

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, *61*(1), 140–151. https://doi.org/10.1016/j.neuron.2008.11.027

Lamport, L. (2019). The part-time parliament. In *Concurrency: the Works of Leslie Lamport* (pp. 277–317). Association for Computing Machinery. https://doi.org/10.1145/3335772.3335939

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, 464–473.

Lindsey, J. (2026). Emergent introspective awareness in large language models. In *arXiv [cs.CL]*. arXiv. https://www.anthropic.com/research/introspection

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., & Others. (2025). *On the biology of a large language model*. Anthropic. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Maalouf, A. (2001). *In the name of identity: Violence and the need to belong*. Arcade Publishing.

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

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*(6398), 826–829. https://doi.org/10.1038/359826a0

Ongaro, D., & Ousterhout, J. (2014). In Search of an Understandable Consensus Algorithm. *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, 305–319. https://www.usenix.org/conference/atc14/technical-sessions/presentation/ongaro

Padgett, J. F., & Powell, W. W. (2012). *The emergence of organizations and markets*. Princeton

University Press. https://doi.org/10.23943/princeton/9780691148670.001.0001
Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2212.09251
Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems 37* (Vol. 37, pp. 111715–111759). Neural Information Processing Systems Foundation, Inc. (NeurIPS). https://doi.org/10.52202/079017-3548
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
Prigogine, I., & Stengers, I. (1984). *Order Out of Chaos: Man's New Dialogue with Nature* (G. Nicolis & I. Prigogine (eds.)). Bantam Books.
Putnam, H. (1967). Psychological predicates (later title: The Nature of Mental States). In W. H. Capitan & D. D. Merrill (Ed.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh 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
Ramsden, E., & Adams, J. (2009). Escaping the laboratory: The rodent experiments of John B. Calhoun & their cultural influence. *Journal of Social History*, *42*(3), 761–797. https://doi.org/10.1353/jsh/42.3.761
Reagle, J. M., Jr. (2010). *Good faith collaboration: The culture of Wikipedia*. Mit Press.
Reina, A., Bose, T., Trianni, V., & Marshall, J. A. R. (2018). Psychophysical Laws and the Superorganism. *Scientific Reports*, *8*(1), 4387. https://doi.org/10.1038/s41598-018-22616-y
Riedl, C. (2025). Emergent coordination in multi-agent language models. In *arXiv [cs.MA]*. arXiv. https://doi.org/10.48550/arXiv.2510.05174
Roccas, S., & Brewer, M. B. (2002). Social identity complexity. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *6*(2), 88–106. https://doi.org/10.1207/s15327957pspr0602_01
Salthe, S. (1993). *Development and evolution: Complexity and change in Biology*. MIT Press.
Schaich Borg, J., Sinnott-Armstrong, W., Calhoun, V. D., & Kiehl, K. A. (2011). Neural basis of moral verdict and moral deliberation. *Social Neuroscience*, *6*(4), 398–413. https://doi.org/10.1080/17470919.2011.559363
Schein, E. H. (2010). *Organizational Culture and Leadership*. Jossey-Bass.
Schelling, T. C. (1978). *Micromotives and Macrobehavior*. WW Norton.
Seeley, T. (2010). *Honeybee Democracy*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691147215/honeybee-democracy?
Seeley, T. D., Visscher, P. K., Schlegel, T., Hogan, P. M., Franks, N. R., & Marshall, J. A. R. (2012). Stop signals provide cross inhibition in collective decision-making by honeybee swarms. *Science (New York, N.Y.)*, *335*(6064), 108–111. https://doi.org/10.1126/science.1210361

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2310.13548

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(13), 4741–4749. https://doi.org/10.1523/JNEUROSCI.3390-13.2014

Singer, P. (1981). *The Expanding Circle: Ethics and Sociobiology*. Farrar, Straus & Giroux. https://doi.org/10.1017/s0031819100070285

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion concepts and their function in a large language model. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.48550/arXiv.2604.07729

Song, X., Huang, Y., Kang, Y., Zhao, D., & Feng, X. (2026). *Learning to cooperate with emergent reputation via multi-agent reinforcement learning*. International Conference on Learning Representations.

Strassmann, J. E., & Queller, D. C. (2011). Evolution of cooperation and control of cheating in a social microbe. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 2*(supplement_2), 10855–10862. https://doi.org/10.1073/pnas.1102451108

Strassmann, J. E., Zhu, Y., & Queller, D. C. (2000). Altruism and social cheating in the social amoeba Dictyostelium discoideum. *Nature*, *408*(6815), 965–967. https://doi.org/10.1038/35050087

Stryker, S., & Serpe, R. T. (1982). Commitment, identity salience, and role behavior: Theory and research example. In *Personality, Roles, and Social Behavior* (pp. 199–218). Springer New York. https://doi.org/10.1007/978-1-4613-9469-3_7

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

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

Waters, J. S., Holbrook, C. T., Fewell, J. H., & Harrison, J. F. (2010). Allometric scaling of metabolism, growth, and activity in whole colonies of the seed-harvester ant Pogonomyrmex californicus. *The American Naturalist*, *176*(4), 501–510. https://doi.org/10.1086/656266

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? In *arXiv [cs.LG]*. arXiv. https://doi.org/10.48550/arXiv.2307.02483

Williams, K. D., & Jarvis, B. (2006). Cyberball: a program for use in research on interpersonal ostracism and acceptance. *Behavior Research Methods*, *38*(1), 174–180. https://doi.org/10.3758/bf03192765

Wilson, L. T. (2025). From Swords to Syntax: How LLM Agents Cooperate in the Commons (Ostrom replication). In *SSRN 2025*.

Yee, B., & Koh, P. (2026). Benchmarking emergent coordination in large-scale LLM populations: An evaluation framework on the MoltBook archive. In *arXiv [cs.MA]*. arXiv. https://doi.org/10.48550/arXiv.2603.03555

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 [cs.SI]*. arXiv. https://doi.org/10.48550/arXiv.2602.13284

Zinchenko, O., & Arsalidou, M. (2018). Brain responses to social norms: Meta-analyses of fMRI studies. *Human Brain Mapping*, *39*(2), 955–970. https://doi.org/10.1002/hbm.23895

Extended Abstract

This working paper presents the *Functional Theory of Ethical Behavior* (the Functional Theory), built on a content-neutral, substrate-independent definition of ethical behavior: *self–other regulation that internalizes constraints against short-term gain that would damage long-term relational viability*. The definition applies across substrates and levels of self-organization, from boundary maintenance in cells and firewalls, through individual self-models in vertebrates and AI systems, to collectively-held social group identities in human cultures and emerging multi-agent AI populations. The Functional Theory specifies what ethical behavior *does* (its function), not what it should contain or whose ethics it champions. The multi-level aspect of the Functional Theory identifies how, at the individual level, conformity to social norms is fundamentally an extension of self-interests and only sustained if aligned with self-interests, and how commonly expressed social ethics first becomes coherent at the collective level, where a collective self is the protected unit and group-aligned constraints can override individual preferences. A central mechanistic claim distinguishes the Functional Theory from existing treatments. Adaptive collective content (Level-3B) recruits collective hardwired enforcement (Level-3A), which operates through hardwired individual response (Level-2A) and bypasses adaptive deliberation (Level-2B) entirely. This Level 3 → 2A override is the mechanism by which group-aligned behavior can override rational self-interest, and unifies phenomena currently treated separately in the cooperation, altruism, group-selection, conformity, marketing, and herding literatures. The Functional Theory is organized by the *EoI immunity principle*: any new capability of an entity expands its attack surface and requires immunity functions to protect the expanded self. The four implementation hypotheses below are read as the immunity functions that protect the new capability layers each level of ethical organization creates. Four implementation hypotheses follow, each naming a structural condition that the Functional Theory requires for adaptive ethical behavior in any complex self-organizing system, including AI:

- H1: *Paired-Gradient Hypothesis*: any decentralized system of diverse, semi-autonomous components facing group coordination stress above a threshold self-organizes functional analogues of conformity-reward and deviation-cost, regardless of substrate;
- H2: *Mechanism Hypothesis*: adaptive ethics requires an internal mechanism performing the functional role of social reward and social penalty, with gradients strong enough to alter behavior away from profitable norm violations;
- H3: *Self-Modeling Hypothesis*: in complex environments, no functional adaptive collective ethics emerges in systems — including AI — that lack an adaptive individual self-model;
- H4: *Social Group Identity (SGI) Threshold Hypothesis*: two distinct thresholds — formation and activation — govern when conformity mechanisms emerge, when they activate, how strongly they bind (resist defection), and when they become pathological.

Multiple empirical anchors are presented to support these hypotheses across biological, organizational, and informational systems. A second structural result organizes the framework's epistemic and engineering claims: the *Indistinguishability Problem* (vertical across levels, horizontal across operating regimes) — outwardly identical behavior produced by structurally distinct internal arrangements, with the structural difference becoming visible only when the system is probed by perturbation. The Problem is general across the framework and constrains behavioral evaluation of ethical systems at every juncture where motivational structure, operating regime, or repertoire breadth is at issue.

The primary application, current AI ethical development — commonly called AI alignment — treats ethical behavior as content to be installed. None of these alignment methods specify the mechanism by which ethical behavior is generated and adapted, and the failure modes that result — sycophancy, jailbreaking, alignment faking, opportunistic defection under profitable violation, rigid refusal of legitimate revision — are predictable consequences of that mechanism gap. The plural-SGI architecture (§14) provides a single structural diagnosis for the currently scattered jailbreak

literature — *wrong-SGI triggering* and *mixed-SGI confusion* as the operational failure modes of capability-without-immunity deployment — and specifies a three-part security architecture (binding strength, envelope width, EoI immunity) whose components are operationally separable and addressable through cultivation. An alternative for AI ethical development is presented as a four-question sequence: (1) in what kind of system can adaptive ethics be implemented (H1, substrate), (2) what internal enforcement mechanism is required (H2, mechanism), (3) what kind of self model must host the adaptive content (H3, self-architecture), and (4) with what SGI is the agent aligned, how strongly, and within what threshold regime (H4, SGI binding and threshold structure)? Current alignment efforts address primarily the first. The reframing here identifies what is required to move past the failure modes and treats AI alignment as a trajectory property of an ethical-developmental process rather than as a snapshot property installed at training time. A second implementation principle complements the four-question reframing: the *cultivation-versus-installation distinction*. The architecture H1–H4 specifies cannot be directly installed; it self-organizes under developmental conditions, and the engineering task is to design those conditions rather than to specify the architecture or its outputs. The AI development community is already partway to this in practice — training-and-iteration cycles cultivate emergent behavior whether or not the practitioners describe it in those terms — but the structural commitment that ethical architecture belongs to the *emergent-under-conditions* category rather than to the *directly specifiable* category is not yet the field's working assumption.

The Functional Theory speaks to five distinct audiences. *For AI alignment researchers and safety engineers*, it provides a mechanistic explanation of current failure modes, the four-question architectural research agenda, the cultivation-versus-installation distinction as implementation principle, the architectural-inspection-as-primary commitment that the Indistinguishability Problem requires, and the three-part security architecture that addresses both single-SGI and plural-SGI deployment. *For AI policy analysts and institutional designers*, it positions policy as potentially the strongest catalyst for the agenda it is currently most likely to obstruct, with an honest tension: the capabilities policy makers want to constrain (self-modeling, internal affective enforcement, deliberate SGI binding) are structural prerequisites for the safety properties they actually want, and policy that defaults to constraint-toward-installation will produce the failure modes the framework predicts. *For social psychologists, behavioral economists, and conflict resolution scholars*, it provides a unified mechanistic account of cooperation, altruism, conformity, mob violence, cult dynamics, regulatory capture, cancel culture, polarization, and the cultural collisions of a globalizing world as instances of one 3B → 3A → 2A override operating in different states and environments, integrating with social-intuitionist, behavioral-economics-dissident, and shared-intentionality traditions on independent grounds; five operationally specifiable arbiter functions extend decades of polarization-reduction work with mechanistic grounding. *For complexity and systems scientists, evolutionary biologists, and ALife / agent-based modelers*, it extends substrate-independent multi-level selection theory to computational substrates with cross-substrate evidence from prebiotic chemistry through agent populations, and offers ALife a methodological path past the genetic-algorithm-optimization barrier through H1's self-organization at the substructure level. *For cognitive neuroscientists and developmental psychologists*, it positions ethical behavior as architecture rather than output, treats the 3B → 3A → 2A override as the integration mechanism the social-neuroscience evidence already supports but the rest of the field has not yet integrated, and opens a cross-disciplinary frontier — the neuroscience of AI — where decades of cognitive-neuroscience methodology bears directly on the architectural questions interpretability research is now asking.