Open Questions In Collective Agency

How Collections Become Agents — And Why We Need to Know

Jonas Hallgren[1]

[1]Equilibria Network

November 2025

Abstract

When does a collection of components become a unified agent with its own goals? This question—simple to ask, profound to answer—cuts across every field studying collective behavior. We can watch it happen: cells organizing into organs, traders forming markets, neurons creating minds, humans building societies. Yet we lack foundational theory to predict when it will happen, how it emerges, or what makes it robust versus fragile. This gap becomes critical as artificial agents join human collectives at scale. We're integrating AI into markets, democracies, and organizations without understanding how hybrid systems discover their own agency. This paper maps the conceptual territory we must navigate, showing why current frameworks fall short and what questions we need to answer. We present not a unified theory but something perhaps more valuable: a clear view of the frontier, showing where different mathematical approaches converge, where they diverge, and what bridges we need to build. This is an invitation to researchers across biology, physics, computer science, neuroscience, and social science—the synthesis we need will emerge from exactly the kind of collective intelligence we're trying to understand.

---

1 The Mystery: What We Can See But Cannot Yet Explain

Picture a salamander that has lost its leg. Within days, cells at the wound site—cells that had settled into stable identities as skin, muscle, or bone—begin to do something remarkable. They abandon their specialized roles and form a structure called a blastema, a mass of apparently undifferentiated tissue. Over the following weeks, this collection of cells reconstructs not just *a* leg, but the *right* leg: properly proportioned to the body, correctly integrated with existing anatomy, complete with bones, muscles, nerves, and blood vessels all in their proper spatial relationships.

Here's what makes this deeply puzzling: no cell in that blastema possesses a blueprint of what a leg looks like. No master controller orchestrates the process. Each cell follows purely local rules—sense chemical gradients, maintain mechanical connections with neighbors, respond to electrical potentials, divide when certain stress thresholds are met. The individual cellular "goals," if we can call them that, are remarkably simple:

- Maintain internal pH within viable ranges

- Follow concentration gradients of specific molecules
- Preserve adhesion strength with neighboring cells
- Minimize mechanical stress through division or reorganization

Yet somehow—through mechanisms we're only beginning to glimpse—these cellular-level behaviors compose into something qualitatively different: a limb-level goal of "reconstruct the missing structure in the anatomically correct location." The salamander solves a staggering coordination problem without anyone—no cell, no subsystem, no homunculus—explicitly representing that problem.

This is what we might call a **growing system**. Not engineered from the top down with explicit specifications, but emerging from the bottom up through local interactions that discover global organization.

1.1 The Same Mystery, Different Scales

Now shift your attention to an ant colony optimizing foraging routes. Individual ants follow laughably simple rules: if you find food, release pheromones on your return path. If you detect pheromones, follow them with probability proportional to concentration, adding your own pheromones if you find food. These local, chemical-gradient-following behaviors—remarkably similar in spirit to what salamander cells do—somehow produce sophisticated network optimization. The colony discovers shortest paths, balances exploration versus exploitation, adapts to changing resource distributions, allocates labor efficiently.

Where is the intelligence? Not in individual ants, whose behavioral repertoire is extremely limited. The colony exhibits collective intelligence that emerges from interactions, not from any ant possessing a global view or strategic plan.

Or consider a prediction market where both humans and AI trading agents bet on future events. In a well-functioning market, prices should aggregate dispersed information—the classic Hayekian insight. Each trader pursues individual profit by trading on private information. Yet through the price mechanism, the market as a whole develops "beliefs" about future events that often exceed any individual trader's knowledge.

But something interesting happens as we introduce more AI agents. As the ratio of algorithmic to human traders shifts, the market's character changes—how quickly it incorporates news, how it responds to rumors, which correlation patterns it amplifies. The system is transforming before our eyes. Yet we cannot answer basic questions: At what point does this human-AI market become a *unified agent* with goals distinct from either humans' intentions or AI's programmed objectives? How do we identify where "the market" ends and "its environment" begins when AI agents simultaneously participate in thousands of markets, carrying information across boundaries?

1.2 The Pattern Across Domains

Table 1 makes the pattern explicit. Across biology, social systems, cognitive systems, and markets, we see the same fundamental puzzle: collections of components with local goals somehow become unified wholes with emergent, qualitatively different goals.

> **Key Insight:** *These aren't separate mysteries requiring separate theories. They're manifestations of a single deep question about how agency scales and transforms across levels of organization. Understanding one might illuminate all the others.*

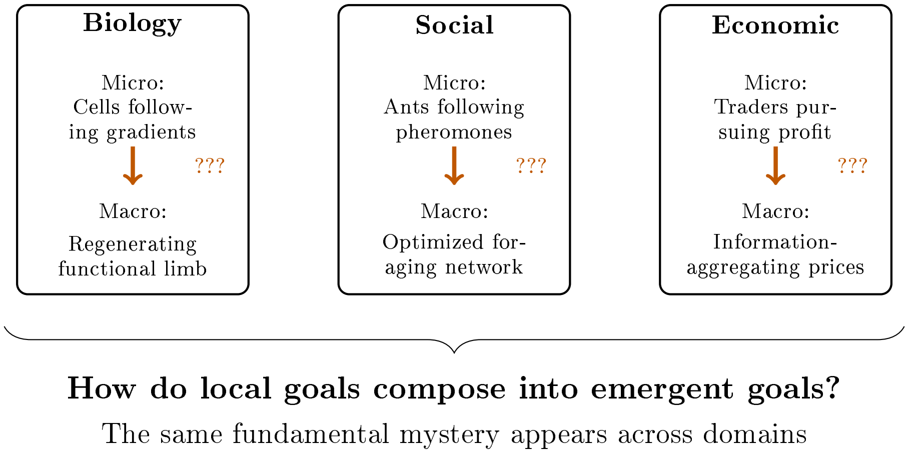
Biology
Micro:
Cells following gradients
???
Macro:
Regenerating functional limb
Social
Micro:
Ants following pheromones
???
Macro:
Optimized foraging network
Economic
Micro:
Traders pursuing profit
???
Macro:
Information-aggregating prices
How do local goals compose into emergent goals?
The same fundamental mystery appears across domains

Figure 1: The central mystery manifests across radically different systems. In each case, simple local behaviors somehow compose into sophisticated collective intelligence. The micro-to-macro transformation is what we need to understand.

Domain	**Components**	**Emergent Collective**	**The Mystery**
Developmental Biology	Cells maintaining homeostasis	Organism with morphological goals	Goal composition
Neural Systems	Neurons with firing thresholds	Unified cognitive agent	Consciousness
Social Insects	Individuals following pheromones	Colony with adaptive strategy	Swarm intelligence
Markets	Self-interested traders	Price-discovering mechanism	Information aggregation
Democracies	Voters with private preferences	Collective decisions	Preference aggregation
Human-AI Systems	Mixed human and AI agents	**???**	**We don't know**

Table 1: The same question appears across domains: how do micro-level behaviors compose into qualitatively different macro-level intelligence? The highlighted row shows why this matters urgently—we're creating hybrid systems without understanding the principles.

1.3 Why This Matters Now

The mystery becomes urgent because we're performing an unprecedented experiment: introducing artificial agents into human collective intelligence systems at scale, in real-time, with enormous stakes.

Consider three scenarios unfolding right now:

Algorithmic Trading in Financial Markets. High-frequency trading algorithms now account for substantial market volume. These AI agents process information faster than humans, detect patterns across correlated assets, and execute complex strategies. But markets evolved as mechanisms for aggregating human information and preferences. What happens to a market's collective intelligence properties when a significant fraction of "agents" are algorithms optimized for profit without human-like information processing, risk preferences, or decision-making constraints? At what ratio of AI to human traders does the market fundamentally change character? Does it become more or less informationally efficient? More or less stable under stress?

We don't know. We're running the experiment without understanding the principles that would let us predict outcomes.

LLMs Mediating Human Discussion. Large language models increasingly facilitate human collective decision-making—summarizing discussions, identifying points of consensus and disagreement, suggesting framings. The LLM becomes part of the communication substrate itself. But human collective intelligence depends on specific information flow architectures. When an AI mediates discussion, it's not just transmitting information neutrally—it's shaping what gets emphasized, what gets filtered, what connections become salient. At what point does the LLM transition from being a tool individuals use to being an agent *within* the collective, influencing the collective's emergent properties?

Algorithmic Content Curation. Recommendation algorithms determine what information reaches which people, shaping attention at civilization scale. These algorithms are themselves evolving through A/B testing and machine learning, being optimized for engagement metrics. The collective intelligence of human society depends on how information flows through social networks. We're fundamentally altering that architecture through algorithms whose effects we barely understand.

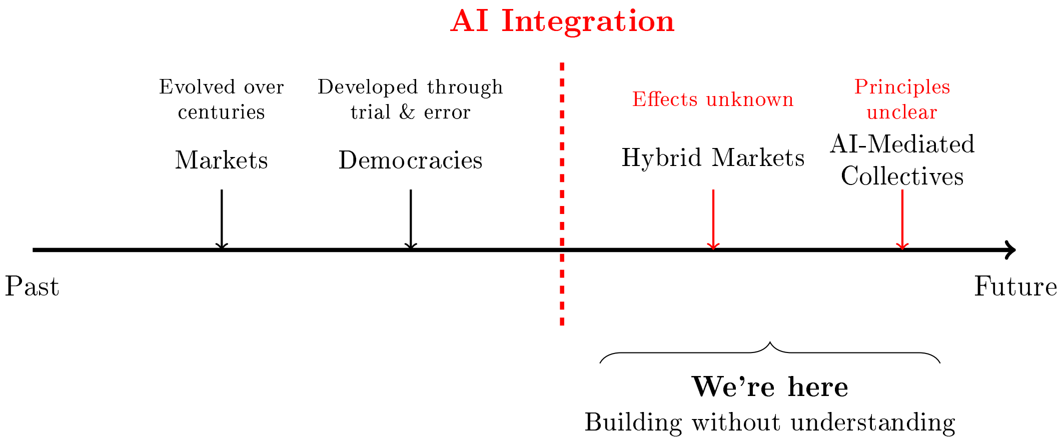

Figure 2: Historical collective intelligence systems evolved gradually, allowing time for adaptation and learning. We're now introducing AI agents at a pace that outstrips our understanding. The gap between deployment and comprehension is the crisis.

This leads us to ask: What exactly is going wrong? Why can't we just extend existing theories to handle these cases?

2 The Diagnosis: Why We're Fundamentally Stuck

The problem runs deeper than you might initially think. It's not that we lack mathematical tools or empirical data. It's that all our dominant frameworks make the same hidden assumption—an assumption that worked fine for engineered systems but breaks down completely for growing collective intelligence.

2.1 The Two Paradigms

There are fundamentally two ways to approach collective systems, and they lead to radically different questions and methods. Figure 3 shows the divergence.

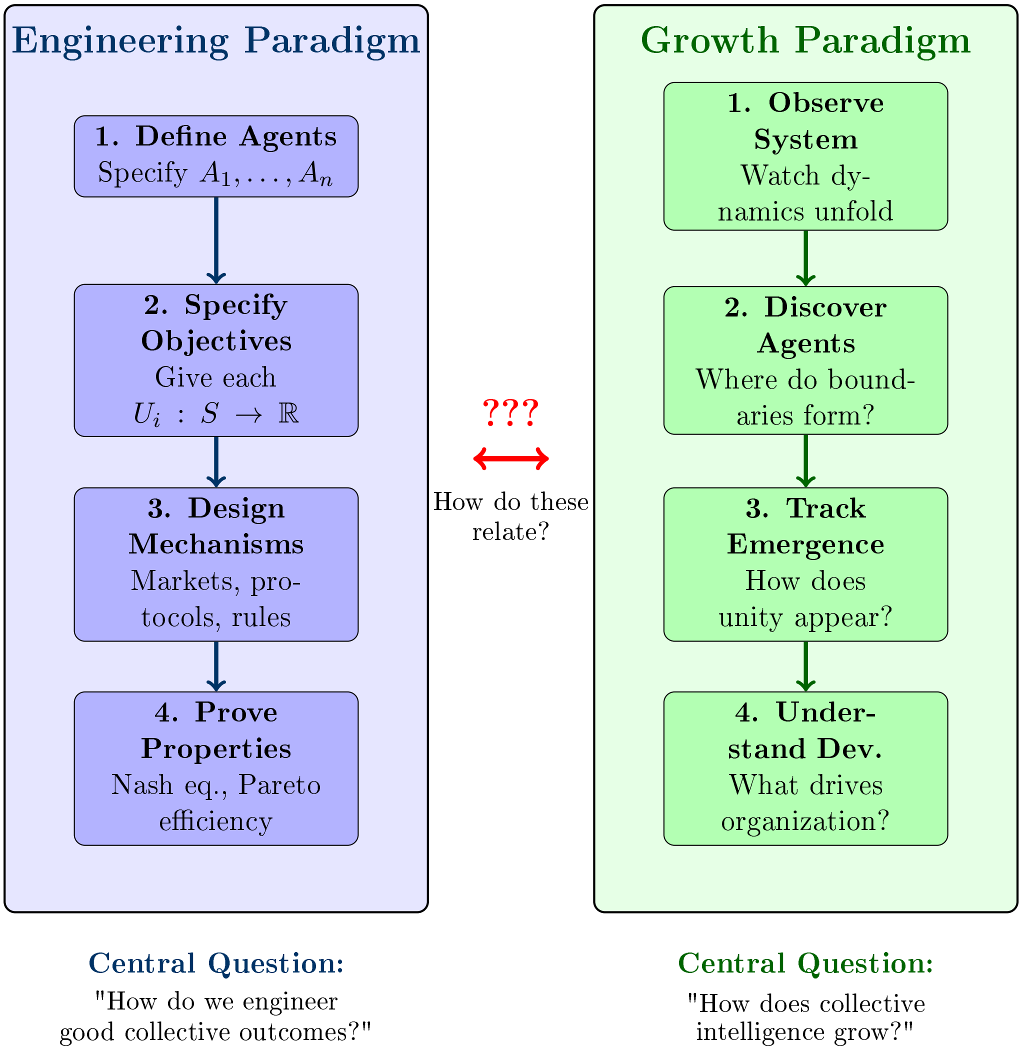

Figure 3: Two fundamentally different approaches to collective systems. The engineering paradigm assumes agents and designs coordination. The growth paradigm observes how agents and coordination emerge. We lack a framework that bridges them.

The Engineering Paradigm dominates computer science, economics, and mechanism design. It asks: "Given agents with these properties, how can we engineer good collective outcomes?" This approach has given us extraordinary tools—game theory, distributed algorithms, mechanism design, multi-agent reinforcement learning. These tools work beautifully when you can specify agents, objectives, and interaction rules up front.

But watch what happens when we try to apply this to our salamander: "Define your agents." Well, are the agents individual cells? Groups of cells? The blastema as a whole? The answer changes depending on what you're trying to predict. "Specify their objectives." Cells maintain homeostasis, but the limb-level goal emerges—it's not programmed into any cell. "Design mechanisms for coordination." The salamander discovered its own coordination mechanisms through evolutionary search, and we're only beginning to understand them.

The engineering paradigm assumes the very things we need to explain: where agents are, what their goals are, how they should coordinate.

The Growth Paradigm appears in developmental biology, evolutionary theory, and complex systems science. It asks: "How do collective intelligence properties emerge from local interactions?" This approach reveals something profound: agents, boundaries, goals, and coordination mechanisms are all *discovered* by the system itself through development.

But this paradigm has its own limitations. It's largely descriptive rather than predictive. It studies what *has* emerged but provides limited guidance for engineering systems that *should* emerge. It lacks the crisp mathematical formulations that make the engineering approach so powerful.

> **Key Insight:** *The crisis we face is this: We're building hybrid human-AI systems using the engineering paradigm (we design AI agents, program their objectives, deploy them into existing mechanisms), but these systems are actually growing systems that will discover their own emergent properties. We're using the wrong conceptual framework for the phenomenon we're creating.*

2.2 The Hidden Assumption

Let me make the problem visceral with a specific example. Consider designing a prediction market. The standard approach:

1. **Define agents:** Traders with beliefs and risk preferences
2. **Specify mechanism:** Continuous double auction with specific rules
3. **Prove properties:** Under certain conditions (common prior, rational expectations), prices converge to true probabilities
4. **Implement:** Build the platform, let people trade

This works remarkably well! Until we start adding AI agents. Now the questions that were assumed away become central:

- **Who are the agents?** Individual traders? Or are some AI agents communicating with each other, effectively forming a larger meta-agent? Do humans using AI trading assistants count as single agents or hybrid systems?
- **What are the boundaries?** Where does "the market" end? If AI agents trade across hundreds of correlated markets simultaneously, are those really separate markets or subcomponents of a larger collective?
- **What are the goals?** Traders pursue profit, but hybrid systems might pursue something else entirely—patterns of correlation that create profit opportunities might become goals in themselves, separate from any human intention.
- **When does unity emerge?** At what point do we have not just traders in a market, but the market-as-agent exhibiting goal-directed behavior (like defending against manipulation or actively seeking out information)?

The engineering paradigm says: "These questions shouldn't matter if we designed the mechanism correctly." But they do matter, because the system is growing and discovering properties we didn't engineer into it.

2.3 Why Current Approaches Fall Short

Let's be specific about what breaks down. Table 2 shows how major research traditions handle (or fail to handle) the central question.

Framework	What It Explains Well	Where It Breaks Down
Game Theory	Strategic interactions between defined agents	Cannot explain agent emergence; assumes rationality and common knowledge
Mechanism Design	How to achieve objectives through rules	Assumes agents and objectives exist; can't design for emergent goals
Multi-Agent RL	How agents learn in interaction	Assumes agent boundaries fixed; struggles with emergent collective behavior
Distributed AI	Coordination protocols for known agents	Cannot handle agents discovering themselves; no theory of boundaries
Complex Systems	General principles of emergence	Often descriptive rather than predictive; lacks engineering applicability
Active Inference	Individual agency through free energy	Single-agent focused; scaling to true collectives unclear

Table 2: Major research frameworks and their limitations. Each powerful within its domain, but none addresses the fundamental question: how do collections become agents?

The pattern is clear: frameworks that enable engineering assume away emergence. Frameworks that study emergence lack engineering precision. We're stuck between two incomplete perspectives, and that's where the hybrid human-AI systems we're building fall through the cracks.

This leads us to ask: What would a better framework look like? What mathematical tools might bridge the gap?

3 The Toolkit: New Lenses for Seeing the Problem

If the engineering paradigm assumes agents and the growth paradigm watches them emerge, we need something that can do both—mathematics precise enough to guide engineering, yet flexible enough to capture emergence. Remarkably, such tools exist. They've been developed in isolation across different fields, each addressing pieces of the puzzle. What we haven't done is understand how they fit together.

This section introduces four mathematical lenses, showing what each reveals and where each falls short. Think of them not as competing theories but as different ways of asking the same underlying question: *Where are the agents?*

3.1 A Master Framework: The View from Multiple Scales

Before diving into specific tools, let's establish the conceptual structure they all share. Figure 4 shows what all these approaches are trying to capture.

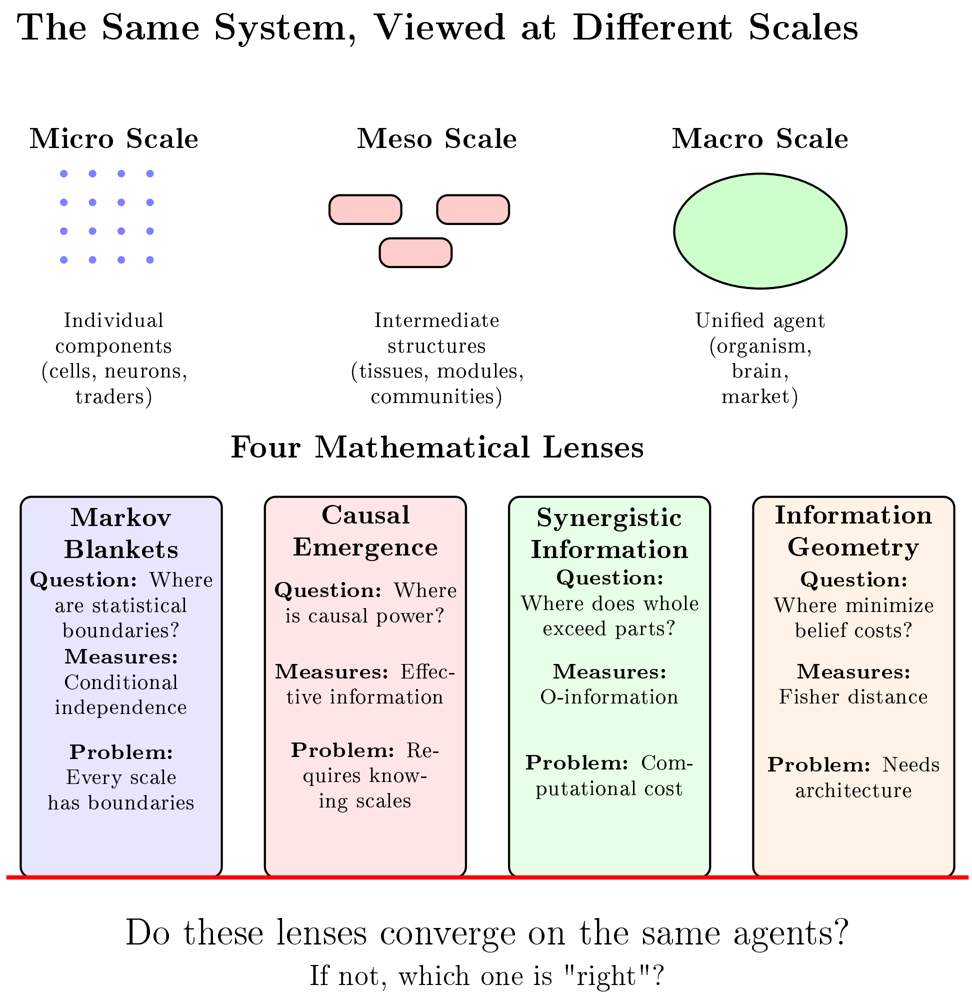
The Same System, Viewed at Different Scales
Micro Scale
Meso Scale
Macro Scale
Individual components (cells, neurons, traders)
Intermediate structures (tissues, modules, communities)
Unified agent (organism, brain, market)
Four Mathematical Lenses
Markov Blankets
Question: Where are statistical boundaries?
Measures: Conditional independence
Problem: Every scale has boundaries
Causal Emergence
Question: Where is causal power?
Measures: Effective information
Problem: Requires knowing scales
Synergistic Information
Question: Where does whole exceed parts?
Measures: O-information
Problem: Computational cost
Information Geometry
Question: Where minimize belief costs?
Measures: Fisher distance
Problem: Needs architecture
Do these lenses converge on the same agents?
If not, which one is "right"?

Figure 4: The master framework: A system can be viewed at multiple scales (micro, meso, macro). Four mathematical lenses offer different ways to identify where agents exist. The central question: Do they agree? If they diverge, what does that tell us?

Now let's examine each lens in detail, understanding both its power and its limitations.

3.2 Lens 1: Markov Blankets and Statistical Boundaries

The active inference framework, developed primarily by Karl Friston and colleagues, offers an elegant mathematical definition of agency that doesn't require assuming preferences, beliefs, or goals from the start. Instead, it derives them from something more fundamental: statistical boundaries.

The Core Idea: An agent is any system with a Markov blanket—a set of states that statistically separate internal states from external states. Think of it as an information boundary: everything the inside "knows" about the outside must come through sensory states, and everything the outside "knows" about the inside must go through active states.

Formally, partition a system's states into:

$$\begin{aligned}
\mu &: \text{Internal states (hidden from environment)}\\
\eta &: \text{External states (hidden from system)}\\
s &: \text{Sensory states (internal depends on these)}\\
a &: \text{Active states (external depends on these)}
\end{aligned}$$

The Markov blanket is (s, a) if:

$$p(\mu, s, a, \eta) = p(\mu|s) \cdot p(s|a, \eta) \cdot p(a|\mu) \cdot p(\eta|a) \tag{1}$$

This means: internal states μ depend only on sensory states s (not directly on external states η), and external states η depend only on active states a (not directly on internal states μ). The blanket (s, a) mediates all influence between inside and outside.

From this purely statistical structure, active inference derives goal-directed behavior: systems act to minimize surprise about their sensory observations, which (through variational inference mathematics) amounts to acting as if they have goals while maintaining beliefs about the world.

What This Lens Reveals: Markov blankets exist everywhere—cell membranes, skin, organizational boundaries, national borders. At every scale where you find conditional independence structure, you can identify a potential agent. This gives us a systematic way to search for agents in complex systems.

Where This Lens Fails: The problem is promiscuity. If Markov blankets exist at every scale, are they all equally "real" agents? Consider a human: there's a Markov blanket at the cell level, at the organ level, at the organism level, potentially at the social level if we consider a person embedded in relationships. Which one is the "real" agent?

The mathematical structure is neutral—it says "here's a statistical boundary" but doesn't tell us which boundaries correspond to genuine agency versus mere descriptive convenience.

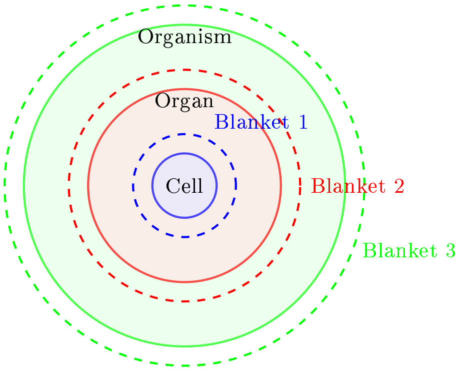
Nested Markov Blankets: Which is the Real Agent?

Figure 5: Markov blankets exist at multiple nested scales. The formalism identifies statistical boundaries but doesn't tell us which boundaries correspond to genuine agency. This is the "promiscuity problem."

3.3 Lens 2: Causal Emergence and Effective Information

Erik Hoel's framework of causal emergence offers a potential resolution to the Markov blanket promiscuity problem. Instead of asking where statistical boundaries are, ask: where does the macro-scale have more causal power than the micro-scale?

The Core Idea: Measure how much a system's current state determines its next state—that's "effective information" (EI). Now compare this at different scales. The agent exists at the scale where EI is maximized—where knowing the macro-state tells you more about the next macro-state than knowing all the micro-details tells you about the next micro-state.

For a Markov chain with transition matrix T:

$$\mathrm{EI}(T) = \sum_{i,j} T_{ij} \log_2 \frac{T_{ij}}{\sum_k T_{ik}} \tag{2}$$

Causal emergence occurs when coarse-graining increases EI:

$$\mathrm{EI}(T_{\text{macro}}) > \mathrm{EI}(T_{\text{micro}}) \tag{3}$$

This happens because coarse-graining can eliminate noise. If micro-states have random fluctuations that don't affect macro-behavior, the micro-description has lower EI (more uncertainty about the next state) while the macro-description has higher EI (less uncertainty about the next macro-state).

What This Lens Reveals: It gives us a principled way to identify the "right" scale. The agent is where causal power concentrates. For a brain, this might be at the level of population dynamics, not individual neurons. For an organism, at the level of organs, not cells. The macro-description captures more of the causal structure.

Where This Lens Fails: Computing EI requires knowing the transition dynamics at every scale you want to compare—the complete Markov chain including all possible states and transitions. For complex systems with many components, this is computationally intractable. Moreover, the lens assumes you already know what scales to look at. How do you identify candidate scales in the first place?

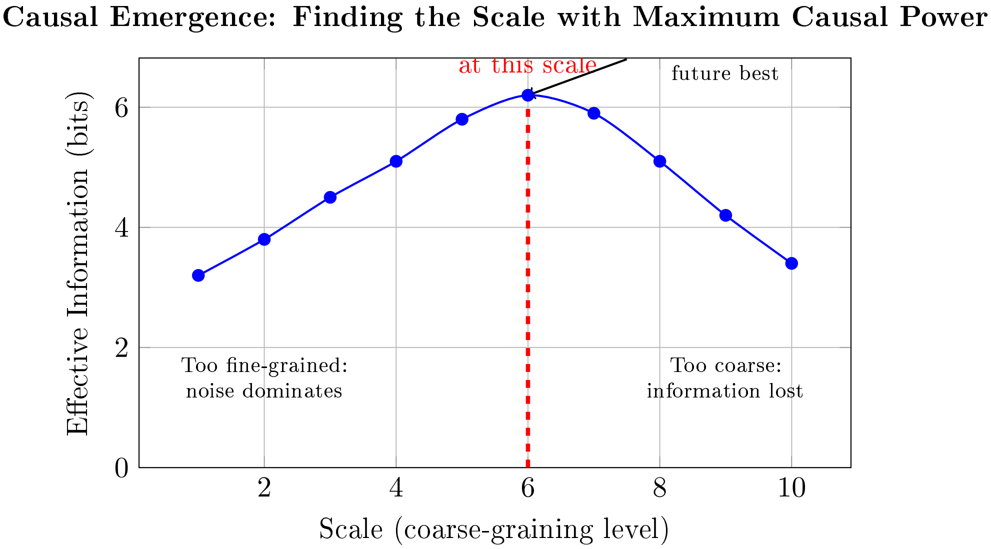

Figure 6: Effective information varies with scale. The agent exists where the macro-description has maximum causal power—more predictive than finer scales (which have noise) or coarser scales (which lose information). This gives a principled answer to "where is the agent?"

3.4 Lens 3: Synergistic Information and Irreducible Integration

Fernando Rosas and colleagues developed a complementary perspective: the agent is where information becomes genuinely synergistic—where the whole contains information that cannot be found in any of its parts.

The Core Idea: Decompose mutual information between a system $\mathbf{X} = \{X_1, \ldots, X_n\}$ and a target Y into three components:

- **Unique information:** In some X_i but not others
- **Redundant information:** Independently available in multiple X_i
- **Synergistic information:** Only accessible by combining X_i—irreducible

The O-information quantifies synergy versus redundancy:

$$\mathcal{O}(\mathbf{X}) = \mathrm{TC}(\mathbf{X}) - \mathrm{DTC}(\mathbf{X}) \tag{4}$$

Negative O-information indicates synergy—the hallmark of emergence. Positive O-information indicates redundancy. Agent boundaries should enclose regions with strong negative O-information.

What This Lens Reveals: Synergistic information is the signature of irreducible wholes. When you measure strong synergy, you've found something that genuinely can't be decomposed—the collective has properties that don't exist in any subset of parts. This is exactly what we mean by a unified agent.

Where This Lens Fails: Computing synergistic information requires estimating high-dimensional probability distributions, which is notoriously difficult with finite data. The O-information grows exponentially complex with system size. For realistic collectives with hundreds or thousands of components, direct computation becomes impossible.

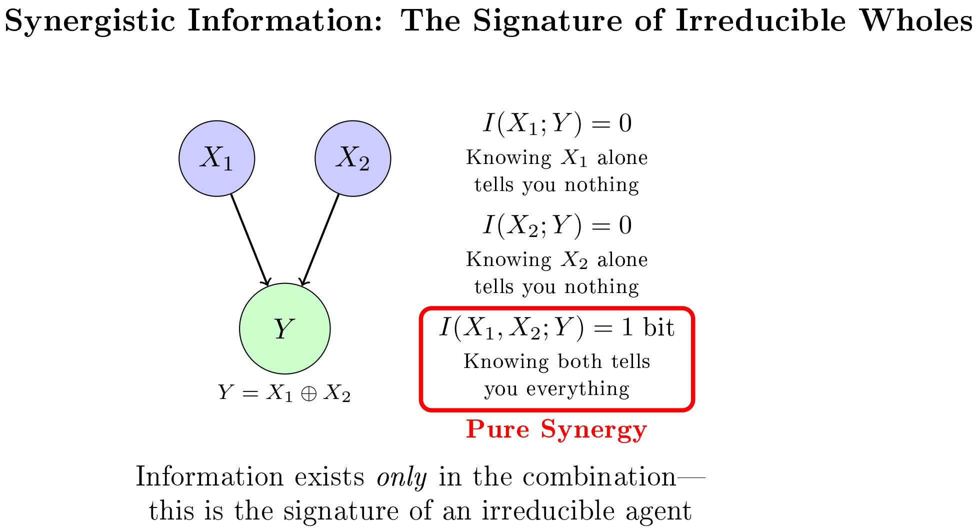
Synergistic Information: The Signature of Irreducible Wholes
X_1
X_2
Y
$Y = X_1 \oplus X_2$
$I(X_1; Y) = 0$
Knowing X_1 alone tells you nothing
$I(X_2; Y) = 0$
Knowing X_2 alone tells you nothing
$I(X_1, X_2; Y) = 1$ bit
Knowing both tells you everything
Pure Synergy
Information exists *only* in the combination— this is the signature of an irreducible agent

Figure 7: XOR gate illustrating pure synergistic information. Neither input alone reveals anything about the output, but together they completely determine it. This is what collective agency looks like mathematically—information that exists only in the whole.

3.5 Lens 4: Information Geometry and Architectural Economics

The fourth lens takes a different approach: instead of asking where agents *are*, ask when they *should* consolidate versus fragment. This framework, developed in the Active Inference Community, uses information geometry to quantify the costs of belief updating.

The Core Idea: There's an economic cost to "changing your mind." In information geometry, beliefs are points in a statistical manifold, and the Fisher information distance measures the cost of updating from belief p to belief q:

$$d_F(p, q) = \sqrt{\sum_i \frac{1}{p(x_i)} [\Delta p(x_i)]^2} \tag{5}$$

Now suppose a system faces environmental novelty ν (how often the world changes requiring belief updates) with model entrenchment λ (how costly updates are). The optimal number of sub-agents minimizes total cost (unreleased paper as of now):

$$K^* = \sqrt{\frac{\lambda \nu d}{C}} \tag{6}$$

where d is typical distance to updated beliefs and C is coordination cost per agent.

What This Lens Reveals: Architecture isn't arbitrary—it's optimized based on fundamental constraints. High novelty environments favor distributed architectures (many sub-agents exploring separately). Stable environments favor centralized architectures (one unified agent). This explains why biological organisms develop different degrees of modularity, why organizations structure differently based on market volatility, why cognitive systems might have dissociable subsystems.

Where This Lens Fails: It requires quantifying λ, ν, and C—the entrenchment, novelty, and coordination costs. For cognitive systems, maybe we can estimate these from neural or computational constraints. But for social organizations or biological development? What are the units? How do we measure across domains? The framework is elegant but demands parameters we often can't access.

3.6 The Central Question: Do These Lenses Converge?

We now have four sophisticated mathematical frameworks, each providing a different answer to "where are the agents?":

- Markov blankets say: wherever there are statistical boundaries
- Causal emergence says: wherever macro-scales have maximum causal power
- Synergistic information says: wherever information is irreducibly integrated
- Information geometry says: wherever architecture minimizes update-coordination trade-offs

Table 3 makes the comparison explicit.

Open Question 1. *When we apply these four frameworks to the same system, do they converge on identifying the same agent boundaries? If they diverge, what does that divergence tell us? Are they measuring different aspects of agency, or are they fundamentally incommensurable?*

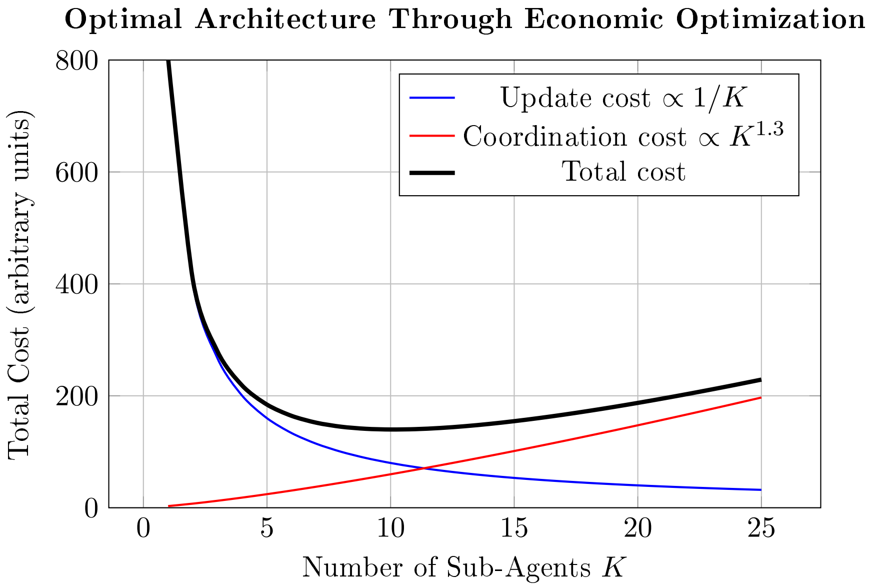

Figure 8: Architecture emerges from balancing update costs (favor distribution) versus coordination costs (favor centralization). The optimal number of sub-agents K^* depends on environmental novelty, model complexity, and interaction costs. This explains why different systems develop different degrees of modularity.

Framework	**Defining Question**	**Measurement**	**Limitation**
Markov Blankets	Where are statistical boundaries?	Conditional independence: $p(\mu \mid \eta, s, a) = p(\mu \mid s)$	Boundaries exist at every scale
Causal Emergence	Where is causal power?	Effective information: $\mathrm{EI}(T_{\text{macro}})$	Requires full transition dynamics
Synergistic Info	Where is information irreducible?	O-information: $\mathcal{O}(\mathbf{X}) < 0$	Exponential computational complexity
Info Geometry	Where minimize costs?	Fisher distance: $d_F(p, q)$	Need to quantify costs across domains

Table 3: Four lenses for identifying agents. Each answers a different question, uses different mathematics, and faces different limitations. The central open question: Do they identify the same agents?

> **Key Insight:** *Agency might not be a simple, unitary property. It might be multi-dimensional, with different aspects (statistical separation, causal efficacy, information integration, architectural optimization) that don't always align. Understanding how they relate could be the key to a unified theory.*

This realization leads us to the heart of the matter: What are the specific open questions we need to answer? Not vague puzzles, but precise, tractable problems that would constitute genuine progress.

4 The Open Questions: A Map of the Frontier

We've seen the mystery (how collections become agents), diagnosed why we're stuck (engineering versus growth paradigms), and examined our mathematical tools (four lenses with different strengths). Now we can finally ask the right questions precisely.

The questions aren't isolated puzzles. They form a connected landscape—a map of conceptual territory where answering one question opens paths to others. Figure 9 shows the structure.

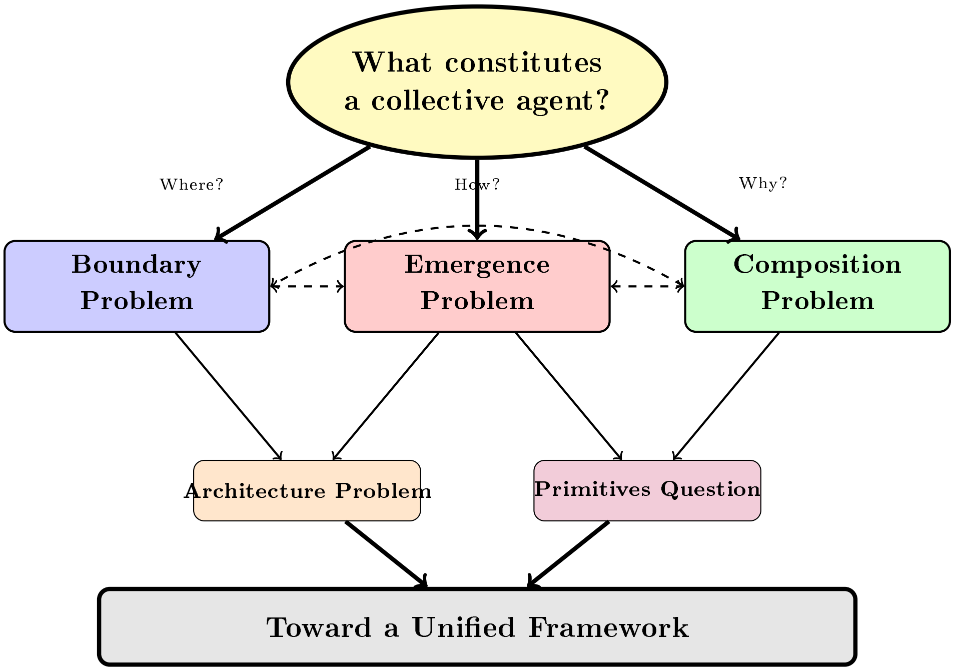

Figure 9: The landscape of open questions. Three primary problems (Boundary, Emergence, Composition) emerge from the central question. They connect to each other and lead to secondary questions about architecture and primitives. All point toward the need for a unified framework.

Let's explore each question cluster, understanding not just what we're asking but why answering it would constitute genuine progress.

4.1 The Boundary Problem: Where Does One Agent End and Another Begin?

The Question: In a growing system where components interact, how do we identify—without assuming the answer—where agent boundaries naturally form, stabilize, or dissolve?

We saw that our four mathematical lenses offer different answers. The boundary problem asks: can we reconcile them? Or if they fundamentally measure different things, can we at least understand the mapping between them?

Why Current Approaches Fail:

- Markov blankets give us too many boundaries (every scale has statistical separation)
- Causal emergence requires knowing dynamics at all scales (computationally intractable)
- Synergistic information faces the curse of dimensionality
- Information geometry presumes we already know the architecture we're trying to derive

What Success Would Look Like:

1. **Computational algorithms** that can identify agent boundaries in complex systems without pre-supposing them. Something that takes raw interaction data and outputs: "these components form a unified agent, these form separate agents, these boundaries are currently forming/dissolving."
2. **Bridging theorems** showing precise relationships between the four frameworks. For example: "Systems with Markov blankets exhibiting synergistic information $> \theta$ will also show causal emergence at the scale where EI is maximized."
3. **Predictive validation:** Apply these methods to developing systems (embryos, organizations, markets) and predict where boundaries will stabilize before it happens. Then check if prediction matches observation.

Open Question 2. *The Inverse Learning Problem: Can we develop algorithms that learn agent boundaries from observations of system dynamics, without assuming those boundaries exist a priori? What's the computational complexity of such inverse learning?*

Open Question 3. *The Reference Frame Problem: Is agency fundamentally observer-dependent, like simultaneity in relativity? If so, what are the transformation rules showing how agent boundaries change between different observational frames? What properties remain invariant across frames?*

4.2 The Emergence Problem: How Does the Whole Become More Than Its Parts?

The Question: What distinguishes a mere collection (cells in a petri dish) from an integrated collective (cells in an organism)? What mathematical properties characterize the transition from collection to agent?

This isn't just asking "when does emergence happen?" but "what *is* emergence in precise, measurable terms?"

Why Current Approaches Fall Short:

- We have multiple frameworks (causal emergence, synergistic information, integrated information) but unclear relationships
- Computational methods don't scale to realistic system sizes
- We lack understanding of dynamics—*how* emergence unfolds over time during development

What Success Would Look Like:

1. **Unified emergence measures** with proven relationships. Something like: "Theorem: Systems with $\Phi > \Phi_c$ necessarily exhibit synergistic information $\mathcal{O} < -\theta$" would show these frameworks capture the same underlying phenomenon.
2. **Developmental trajectories:** Track how emergence measures change during actual developmental processes. Can we identify the critical transition point where a collection becomes a collective? Are there universal scaling laws near these transitions?
3. **Design principles:** From understanding emergence, derive principles for engineering systems with desired emergent properties. "If you want X level of collective agency, you need Y level of integration with Z connectivity structure."

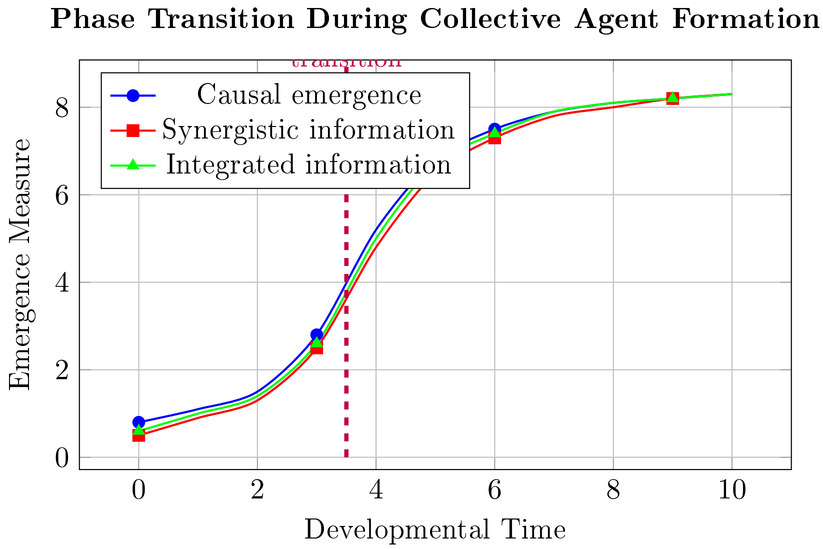

Figure 10: Hypothetical developmental trajectory showing three emergence measures during collective agent formation. All three show rapid increase around the same critical transition point. Question: Do they always align like this, or can they diverge? What causes the transition?

Open Question 4. *The Phase Transition Question: Are there universal properties near the critical point where collections become collectives? Power laws, critical exponents, scaling behaviors? Can we predict the transition point from early observations?*

Open Question 5. *The Measure Convergence Question: Do causal emergence, synergistic information, and integrated information always identify the same systems as "emerged"? If they diverge, what does that divergence reveal about different aspects of collective agency?*

4.3 The Composition Problem: How Do Goals Scale Across Levels?

The Question: How do simple, local micro-goals compose into complex, qualitatively different macro-goals? Not just how goals coordinate, but how genuinely new intentionality emerges at larger scales.

This is perhaps the deepest puzzle. When salamander cells, each pursuing homeostasis, collectively build a limb, where does the limb-level goal come from? It's not programmed into any cell. It's not the sum of cellular goals. It's emergent—but through what mechanism?

Why Current Approaches Fall Short:

- Hierarchical active inference assumes higher levels provide priors, but where does the top-level prior originate?
- Category theory provides elegant composition rules, but risks just renaming the problem with fancier notation
- Information geometry shows how constraints combine, but doesn't explain how combined constraints become *less* restrictive (allowing more complex behavior) at macro-scales

What Success Would Look Like:

1. **Formal framework** showing how goal representations transform across scales. Not just "micro-goals plus interaction structure yields macro-goals" but *what* about the interaction structure enables qualitative transformation.

2. **Predictive power:** Given micro-goals and interaction topology, predict what macro-goals will emerge. Test this on biological development (can we predict morphological targets from cellular behavior rules?) and social systems (can we predict organizational goals from individual incentives plus structure?).

3. **Engineering applications:** Design interaction structures that reliably produce desired macro-goals from given micro-goals. This would enable genuine bottom-up design of collective intelligence systems.

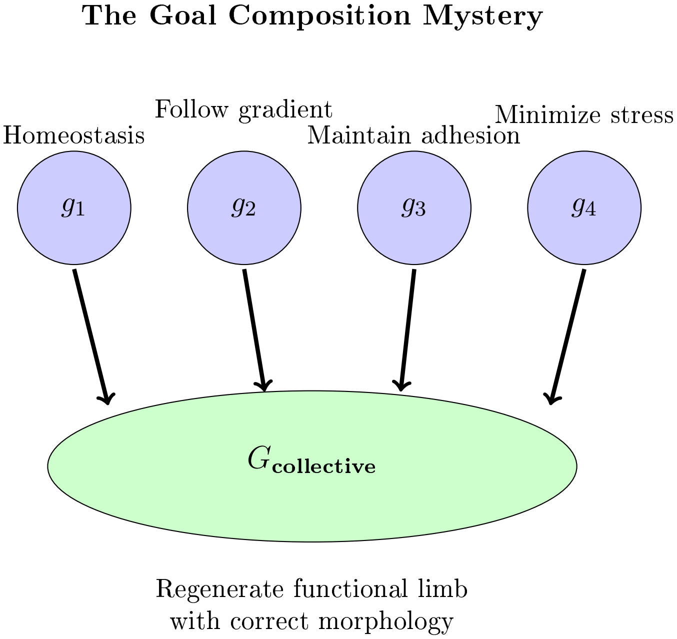

Figure 11: The goal composition problem. Cellular goals are local optimization problems. The limb-level goal involves complex spatial structures that no cell represents. How does this transformation happen?

Open Question 6. *The Relational Goals Question: Michael Levin's work suggests goals are relational—determined by context rather than intrinsic. How do we formalize systems where goals emerge from the interaction between internal states and collective patterns, where those patterns are themselves created by goal-directed behavior? What makes such circular systems stable?*

Open Question 7. *The Scaling Question: As goals compose across scales, what mathematical properties are preserved versus transformed? Is there something analogous to renormalization group transformations showing how goal representations change but certain invariants remain?*

4.4 The Architecture Problem: When Should Systems Consolidate?

The Question: Given a task environment, what determines optimal collective architecture? When should a system consolidate into a unified agent versus remaining distributed? How many sub-agents should exist and how should they interact?

This flips the perspective: rather than asking how collective agency emerges, ask when it *should* emerge. This connects to engineering—if we understand architectural principles, we can design better collectives.

Why Current Approaches Fall Short:

- Information geometry framework is elegant but requires quantifying costs (λ, ν, C) across domains
- Limited to cognitive systems; unclear how to extend to biological development or social organization
- Assumes a single optimization criterion (minimize total cost) but real systems face multiple, potentially conflicting objectives

What Success Would Look Like:

1. **Phase diagrams** showing which architectures are optimal in which regions of parameter space (environmental novelty, coordination costs, task complexity, etc.). This would let us predict architectural transitions before they happen.
2. **Cross-domain validation:** Does the same optimization principle explain modularity in neural systems, biological organisms, organizations, and markets? If so, what are the universal parameters? If not, what varies?
3. **Design tools:** Given environmental parameters and desired performance, compute optimal architecture. Use this for organizational design, AI system architecture, institutional structure.

Environment	**Novelty Rate**	**Optimal K^***	**Architecture**
Highly stable	Very low	≈ 1	Centralized
Slowly changing	Low	$2-3$	Simple hierarchy
Moderately dynamic	Medium	$5-10$	Modular
Rapidly fluctuating	High	$20-50$	Distributed
Hyper-volatile	Very high	100+	Swarm-like

Table 4: Predicted optimal architectures based on environmental novelty. Higher novelty favors more distributed structures. Can we validate these predictions across biological, cognitive, and social systems?

Open Question 8. *The Symmetry-Breaking Question: Can we formalize architecture emergence through symmetry groups? Different collective structures (markets, hierarchies, networks) correspond to different symmetry-breaking patterns. What determines which symmetries break first? Are there conservation laws that persist across architectural transitions?*

Open Question 9. *The Multi-Objective Question: Real systems optimize for multiple objectives simultaneously (speed, accuracy, robustness, adaptability). How does multi-objective optimization change architectural predictions? Are there Pareto frontiers in architecture space?*

4.5 The Primitives Question: What Are the Building Blocks?

The Question: What are the fundamental elements from which a theory of collective intelligence should be built? Agents? Processes? Information flows? Constraints? This is the most foundational question of all.

Different choices of primitives lead to radically different theories. Multi-agent systems take agents as primitive. Active inference takes free energy minimization as primitive. Category theory takes processes as primitive. Each choice opens certain insights while foreclosing others.

What Success Would Look Like:

1. **Comparative analysis** showing what each choice of primitives enables and constrains. Not arguing one is "right," but mapping the space of theoretical possibilities.

2. **Bridging theorems** showing how different primitive choices relate. "Theorem: A process-based formulation with these properties is equivalent to an agent-based formulation with these properties under these conditions."

3. **Principled choice criteria:** What determines which primitives are appropriate for which domains or questions? Can we develop meta-theory about theory choice?

Open Question 10. *The Process vs. Object Question: Should collective intelligence theory take processes/transformations as primitive (category-theoretic approach) or states/agents (set-theoretic approach)? What can each formulation express that the other cannot? Where do they converge?*

Open Question 11. *The Langlands Question: Is there a "Langlands program for collective intelligence"—deep structural connections between seemingly disparate mathematical approaches (information theory, causal emergence, active inference, category theory) that would reveal they're all describing the same underlying reality from different angles?*

5 The Invitation: How You Can Contribute

If you've read this far, you've seen both the depth of the challenge and the tangibility of progress pathways. These aren't vague philosophical puzzles—they're precise mathematical questions with clear success criteria. But solving them requires expertise from multiple fields working together.

5.1 For Biologists: You Have the Existence Proofs

You've watched collective intelligence emerge in developing embryos, evolving populations, regenerating tissues for your entire career. You know it's possible. You've seen cells discover organization, organisms adapt to novel environments, ecosystems maintain homeostasis.

What you can contribute:

- Help us formalize what you observe. When you say "the tissue knows what it should become," what information flows make that possible?

- Provide experimental systems for testing theoretical predictions. If we predict boundary formation should happen at time t under conditions C, can you design experiments to check?

- Challenge our frameworks with biological complexity. Where do our neat mathematical models break down when confronted with actual developmental biology?

Specific research directions:

1. Apply emergence measures (causal emergence, synergistic information) to developing systems at multiple time points. Track how emergence unfolds.
2. Test architectural predictions: Do organisms in volatile environments evolve more modular architectures, as information geometry predicts?
3. Map the "goal composition" process: How do bioelectric patterns and morphogen gradients encode limb-level goals that no individual cell possesses?

5.2 For Computer Scientists: Build the Tools

You create algorithms, design systems, implement mathematics in silicon. You can take these abstract frameworks and make them computational—both testing their validity and enabling practical applications.

What you can contribute:

- Develop scalable algorithms for computing emergence measures (EI, O-information, Φ) that work on realistic system sizes
- Implement boundary discovery algorithms that don't assume agent structure
- Create simulation frameworks for testing how hybrid human-AI systems develop collective properties

Specific research directions:

1. Build multi-agent simulation environments where agents can discover their own boundaries through interaction. Do emergent boundaries match theoretical predictions?
2. Implement and compare all four mathematical lenses (Markov blankets, causal emergence, synergistic information, information geometry) on the same systems. Where do they converge? Where diverge?
3. Create visualization tools that make collective intelligence dynamics comprehensible—show how information flows, where boundaries form, when emergence happens.

5.3 For Physicists: Find the Universal Structure

You've discovered universal principles in phase transitions, renormalization group flows, critical phenomena. You see through surface details to underlying mathematical structure. That's exactly what collective intelligence needs.

What you can contribute:

- Identify universal properties near critical transitions where collections become collectives. Power laws? Scaling exponents? Critical phenomena?
- Develop renormalization group approaches for coarse-graining collective systems while preserving essential dynamics
- Find conservation laws or invariants that persist across architectural transitions or scale changes

Specific research directions:

1. Map the phase diagram of collective intelligence: What regions of parameter space (coupling strength, system size, interaction range) support different collective behaviors?

2. Develop statistical mechanics approaches to collective agency, treating agents as degrees of freedom that can order/disorder through phase transitions.

3. Find the "thermodynamic" principles—what quantities are conserved or optimized during collective intelligence processes?

5.4 For Neuroscientists: Scale from Neurons to Minds

You study how neural collectives become unified minds every day. You understand both the microscopic (neural) and macroscopic (cognitive) scales, and crucially, you've developed frameworks (like active inference) that bridge them.

What you can contribute:

- Test whether principles discovered in neural systems generalize to other collectives. Is the brain a special case or an example of universal principles?

- Develop hierarchical active inference models that scale beyond individual agents to genuine collectives

- Provide empirical measures: We can measure neural activity at many scales simultaneously. What do causal emergence or synergistic information look like in actual brains?

Specific research directions:

1. Measure effective information at multiple scales (neurons, columns, regions, whole-brain) during cognitive tasks. Where is it maximized? Does that match where we intuitively locate "the agent"?

2. Test architectural predictions: Do brains in high-novelty environments (e.g., during learning) show more distributed activity, as information geometry predicts?

3. Develop computational models where neural populations discover their own organization through free energy minimization, without pre-specified architecture.

5.5 For Social Scientists: Test in the Wild

You observe human collectives organizing themselves in markets, democracies, organizations, social movements. You see collective intelligence succeeding and failing in the real world, with real stakes.

What you can contribute:

- Provide naturalistic test cases for theoretical predictions. When do markets exhibit collective agency? When do organizations discover new goals beyond those of their members?

- Challenge our frameworks with social complexity. Human collectives involve culture, norms, institutions—aspects our mathematics might be missing.

- Identify where hybrid human-AI systems are already emerging. Document how they differ from purely human collectives.

Specific research directions:

1. Study prediction markets as AI agent participation increases. Can we measure when the market transitions from "humans trading with AI tools" to "hybrid collective with emergent properties"?
2. Track organizational development: Apply boundary detection algorithms to communication networks in organizations. Do discovered boundaries match formal structure?
3. Analyze democratic systems: How does information flow architecture affect collective decision quality? Can we optimize structure to enhance collective intelligence?

5.6 For Mathematicians: Build the Bridges

You see deep structure. You prove theorems showing how apparently different mathematical objects are actually the same thing viewed differently. That's exactly what we need to connect these frameworks.

What you can contribute:

- Prove the bridging theorems showing how information theory, active inference, causal emergence, and category theory relate
- Develop the formal machinery needed to make these connections rigorous
- Identify what's universal versus domain-specific in collective intelligence mathematics

Specific research directions:

1. Formalize the relationship between Markov blankets and causal boundaries. Theorem: "Systems with Markov blankets satisfying property X exhibit causal emergence at scale Y."
2. Develop category-theoretic frameworks that subsume both process-based and agent-based approaches, showing them as different presentations of the same underlying structure.
3. Create information-geometric formulations of goal composition, showing how constraint manifolds combine to produce emergent goal structures.