A Taxonomy of Agents From The Intentional Stance: Predictive Power, Compression, and Mathematical Lenses

Jonas Hallgren

February 27, 2026

Abstract

This paper reframes the interdisciplinary study of agency not as a search for inherent agent properties, but as an exploration of effective modeling strategies. Drawing inspiration from Dennett's intentional stance, we argue that attributing agency features like memory or theory of mind is a form of cognitive compression, chosen for its predictive power in specific contexts. We first look at the trajectory of mathematical theories related to agency modelling in relevant fields like AI/Robotics, Cybernetics, Cognitive Science, Economics and Biology. We then use these models to distinguish how the different fields model agents and through this find a meta framework for viewing agents based on first principle mathematics. This is a taxonomic description of what has been done rather than an absolute description of what possible theories of agents are. Through this work, we hope to create a lot more specificity in how people use the word "agent" as it is woefully under-specified.

1 Introduction: When Different Scientists See Different Agents

A slime mold placed at one end of a maze will find the shortest path to food at the other end, often outperforming sophisticated computer algorithms. It has no brain, no nervous system, no centralized control. Yet it explores multiple routes simultaneously, abandons unsuccessful paths, and optimizes its network for efficient resource distribution. When researchers cut the optimal path, it quickly discovers the second-best route.

Here is the puzzle: Is this agent-like behavior, or just physics?

Your answer reveals more about your scientific training than about the slime mold itself. A control theorist sees a feedback system maintaining homeostatic goals through error correction—minimal agency requiring only setpoints and negative feedback loops [Powers, 1973, 1978, Heylighen, 2022]. A cognitive scientist might deny agency entirely, noting the absence of memory systems, explicit goal representation, or theory of mind [Cooper and Peebles, 2015, Sowa, 2011]. A developmental biologist observes clear goal-directed behavior and adaptive resource allocation, treating it as a paradigmatic example of distributed intelligence [Levin, 2023, Pezzulo and Levin, 2016]. A computer scientist sees emergent optimization without "true" agency—distributed processing through physical substrates [Masterman et al., 2024].

These are not merely different perspectives on the same phenomenon. Each field would design different experiments, make different predictions, and propose different interventions. The control theorist predicts behavior through differential equations governing chemical gradients [Mulder et al., 2018]. The biologist models adaptive responses to environmental changes through multiscale competency architectures [Manicka and Levin, 2022]. The computer scientist analyzes algorithmic complexity of the parallel search process. They are using fundamentally different mathematical frameworks to compress the same observable behavior.

Consider a second example that reveals similar patterns. When bacteria develop antibiotic resistance, they exhibit sophisticated collective strategies: some cells sacrifice themselves to protect the colony, others develop costly resistance mechanisms they share with neighbors, still others act as "scouts" testing environmental dangers. This resembles military tactics—reconnaissance, sacrifice for group benefit, resource sharing, coordinated defense. Are bacterial colonies strategic agents? They seem to outmaneuver human medical interventions through what appears to be planning, cooperation, and adaptive learning.

Evolutionary biologists model these dynamics using minimal agency—simple reactive strategies with payoff-driven imitation and myopic updating, sufficient to generate evolutionarily stable strategies through

replicator dynamics [Hofbauer and Sigmund, 1998, Broom and Cannings, 2010]. The mathematical framework treats agents as frequency-dependent learners responding to selection pressures without requiring explicit strategic reasoning [Aoki and Feldman, 2014, Nowak et al., 2010]. Behavioral economists might instead apply level-k thinking models with bounded strategic reasoning [Camerer et al., 2004, Crawford et al., 2013]. Cognitive scientists would likely reject attributions of planning or cooperation, noting the absence of mental representation [Doumas and Hummel, 2012]. Each compression strategy enables different predictions and interventions.

A third example brings these challenges into urgent relief: large language models like GPT-4. When we describe these systems as "agents" or discuss their "agentic behavior" [Park et al., 2023], we invoke agency concepts without clear foundations for what this means. Computer scientists debate whether they implement strategic reasoning through chain-of-thought prompting or merely pattern-match at scale [Masterman et al., 2024]. Cognitive scientists analyze whether they possess genuine memory systems or simply maintain context windows [Kriegeskorte and Douglas, 2018]. Economists examine whether they exhibit bounded rationality in decision-making scenarios [Camerer et al., 2003]. The rapid deployment of such systems in critical applications—from financial trading to medical diagnosis—proceeds despite fundamental uncertainty about which agency concepts apply and why [Russell, 2019, Amodei et al., 2016].

1.1 The Core Phenomenon: Systematic Differences in Agency Modeling

These examples illustrate a broader pattern that has become increasingly problematic as AI systems proliferate and scientific domains become more interdisciplinary. When researchers across fields study systems exhibiting apparently purposeful behavior, they develop strikingly different answers to fundamental questions: Which features are necessary for adequate modeling versus merely sufficient? When does memory become essential rather than optional? What role does strategic reasoning play in explaining observed behavior? How do we model goal-directedness without anthropomorphizing?

These are not merely terminological disagreements. Different fields make different mathematical commitments that lead to different predictions, different experimental designs, and different practical interventions. The behavioral economist's level-k thinking model [Camerer et al., 2004] predicts systematically different outcomes than the evolutionary biologist's replicator dynamics [Hofbauer and Sigmund, 1998], even when both address strategic behavior in populations. The control theorist's feedback-based framework [Powers, 1978, Carver and Scheier, 2002] generates different testable predictions than the developmental biologist's multiscale homeostatic architecture [Pio-Lopez et al., 2023], despite both explaining goal-directed behavior in biological systems.

1.2 Why This Matters Now

Three developments make understanding these differences urgent. First, AI system deployment creates practical stakes for theoretical precision. As we deploy increasingly sophisticated AI systems described as "autonomous agents," our theoretical confusion about agency translates directly into practical uncertainty about system behavior, safety, and alignment [Amodei et al., 2016, Russell, 2019]. When AI safety researchers discuss "agent foundations," they typically envision utility maximizers updating beliefs through Bayes' rule [Demski, 2018, Wentworth, 2022]. Yet developmental biologists studying goal-directed cellular collectives use entirely different mathematical frameworks [Levin, 2012, 2023]. If we are building artificial agents, which framework should guide our design?

Second, interdisciplinary research on modern challenges—from pandemic response to climate modeling to technological governance—requires combining insights across disciplines. Yet researchers trained in different traditions literally see different things when examining the same systems. Without frameworks for translation, valuable insights remain trapped within disciplinary silos. Behavioral economics has successfully bridged psychology and economics through bounded rationality models [Kahneman, 2011, Simon, 1982], and cognitive science has developed architectural integration across neural and behavioral levels [Anderson, 2007, Cooper and Peebles, 2015]. However, no systematic framework exists for comparing agency modeling approaches across the full range of relevant scientific domains.

Third, existing theoretical foundations remain incomplete despite decades of agent foundations research in AI and philosophy. Work has largely developed in isolation: behavioral economics synthesizes experimental

findings within game-theoretic frameworks [Camerer, 1997, Crawford et al., 2013], cognitive science attempts unified computational architectures [Sowa, 2011, Doumas and Hummel, 2012], and AI alignment develops mathematical tools for agent boundaries and optimal abstractions [Demski, 2018, Flint, 2020, Wentworth, 2022]. These approaches remain disconnected, lacking unified foundations for comparing their compression strategies and predictive trade-offs.

1.3 Our Approach: Systematic Cross-Domain Analysis

Rather than proposing yet another definition of agency or attempting theoretical unification, we take a different approach: we systematically analyze how established scientific disciplines have already solved agency modeling problems in their respective domains. We treat the diversity of agency concepts across fields not as confusion to resolve, but as data revealing how different predictive challenges and observational constraints lead to different optimal modeling strategies.

This perspective draws inspiration from Dennett's "intentional stance" [Dennett, 1987], which treats agency attribution as a predictive strategy chosen for its effectiveness in specific contexts. We extend this insight through information-theoretic analysis, viewing agency features as components of models that optimize the trade-off between descriptive complexity and predictive accuracy [Rissanen, 1978, Cover and Thomas, 2006]. Just as physicists choose kinetic theory versus thermodynamics depending on the scale of analysis, researchers across disciplines have developed different agency models optimized for their unique predictive challenges and observational constraints.

Our contribution is threefold. First, we provide empirical documentation through systematic literature review across six major domains—behavioral economics, evolutionary systems, developmental biology, AI/robotics, control theory, and cognitive science. We document which agency features each field treats as necessary versus sufficient for adequate prediction. Table 1 presents this analysis, revealing systematic patterns beneath apparent disagreements.

Second, we develop an explanatory framework showing that these patterns reflect optimization under different constraints. Control theory evolved minimal agency concepts because engineers needed reliable goal-seeking with computational efficiency [Powers, 1973, Mulder et al., 2018]. Cognitive science developed complex memory architectures because human behavior exhibits sophisticated temporal integration that simpler models cannot capture [Cooper and Peebles, 2015, Sowa, 2011]. Evolutionary theory demonstrates how sophisticated population-level outcomes emerge from minimal individual agency through selection pressures [Hofbauer and Sigmund, 1998, Aoki and Feldman, 2014]. These are not arbitrary choices—they represent adaptive responses to domain-specific modeling challenges.

Third, we demonstrate practical implications for AI system design and interdisciplinary research. Understanding when different modeling approaches work enables more principled system design. When building AI systems, we can systematically analyze which agency features become necessary given our specific environment and constraints, rather than defaulting to familiar approaches. Recent work on AI agent architectures [Masterman et al., 2024] and multi-agent learning [Li et al., 2022] would benefit from explicit frameworks for comparing different agency modeling strategies.

1.4 Structure and Scope

Section 2 develops three detailed examples showing exactly how different fields model specific systems, including the actual mathematics and competing predictions. Section 3 presents our systematic cross-domain analysis, documenting necessity/sufficiency patterns for major agency features across behavioral economics [Camerer et al., 2003, Crawford et al., 2013], evolutionary systems [Hofbauer and Sigmund, 1998, Broom and Cannings, 2010], developmental biology [Levin, 2023, Pezzulo and Levin, 2016], AI/robotics [Masterman et al., 2024, Vernon et al., 2007], control theory [Powers, 1978, Heylighen, 2022], and cognitive science [Cooper and Peebles, 2015, Kriegeskorte and Douglas, 2018]. Section 4 analyzes the logical structure underlying these patterns, revealing hierarchical dependencies between capabilities. Section 5 discusses implications for AI system design and interdisciplinary research.

We make no claim to have solved "the problem of agency." Instead, we demonstrate that agency modeling across successful scientific domains follows systematic patterns reflecting computational principles rather than philosophical commitments. By making these patterns explicit, we provide infrastructure for more

principled approaches to understanding and designing intelligent systems. This work complements recent efforts in AI alignment [Demski, 2018, Wentworth, 2022] by providing empirical grounding from established scientific disciplines that have successfully navigated similar agency modeling challenges.

2 Three Case Studies: The Same System Through Different Lenses

To make concrete how different fields compress agency into different mathematical frameworks, we examine three specific systems in detail. For each, we show the actual modeling approaches, predictions, and trade-offs that different domains make. These case studies reveal systematic patterns in how different predictive goals and observational constraints lead to different optimal compressions.

2.1 The Slime Mold Pathfinding Problem

2.1.1 The Phenomenon

Physarum polycephalum, a single-celled organism, demonstrates remarkable problem-solving behavior documented extensively in experimental studies. Placed in a maze with food sources at multiple locations, it explores the full maze space initially, sending protoplasmic tubes down all available paths. It gradually retracts from unsuccessful branches while reinforcing successful routes, converging on the shortest path connecting food sources. It redistributes resources efficiently across the network and adapts quickly when the environment changes. This behavior has inspired practical algorithms for network optimization problems. The question is how to model the agency underlying this behavior.

2.1.2 Control Theory: Minimal Agency Through Feedback

Control theorists model this system as a network of coupled feedback loops maintaining homeostatic goals [Powers, 1973, Heylighen, 2022]. The mathematical framework requires only state variables, reference signals, and feedback dynamics. Let $\rho_i(t)$ represent the protoplasm density in tube segment i at time t. Each tube segment maintains a setpoint concentration ρ_i^* related to nutrient availability. The system evolves according to differential equations relating current state to error signals and diffusion between connected segments:

$$\frac{d\rho_i}{dt} = \alpha(\rho_i^* - \rho_i) + \beta \sum_{j \in N(i)} (\rho_j - \rho_i) \tag{1}$$

where α controls local feedback strength and β governs diffusion between connected segments. This minimal model predicts rapid convergence to optimal paths through simple error minimization, requires no explicit memory beyond current state, exhibits robustness through automatic adaptation to perturbations, and scales to arbitrary network complexity without additional mechanisms [Mulder et al., 2018, Carver and Scheier, 2002].

The control-theoretic compression treats strategic reasoning, memory systems, and goal representation as unnecessary for explaining observed behavior. Purposiveness emerges entirely from the feedback architecture. This approach reflects the field's emphasis on computational parsimony and mathematical tractability [Toates and Archer, 1978, Frank, 2018], prioritizing models that achieve reliable goal-seeking with minimal computational requirements.

2.1.3 Cognitive Science: Rejecting Agency Attribution

Cognitive scientists examining the same system would likely reject agency attribution entirely [Cooper and Peebles, 2015, Doumas and Hummel, 2012]. Their compression strategy emphasizes the absence of core cognitive features considered necessary for genuine agency. The system lacks working memory for maintaining task-relevant information, long-term memory for learning from experience, mental representation of goals or environmental structure, strategic reasoning about alternative paths, and theory of mind (obviously inapplicable in this context) [Sowa, 2011, Kriegeskorte and Douglas, 2018].

From this perspective, describing the slime mold as "solving" the maze or "choosing" optimal paths constitutes anthropomorphization. The behavior results from local chemical gradients and mechanical constraints, not from information processing in any meaningful cognitive sense. This framework predicts inability to learn improved strategies from experience, no transfer of "knowledge" between different maze configurations, purely reactive responses without temporal integration, and performance determined entirely by current chemical concentrations.

The cognitive science framework sets a high bar for agency attribution, treating sophisticated memory and representational systems as necessary features that single-celled organisms clearly lack [Fitch, 2014, Cooper and Peebles, 2015]. This reflects the field's commitment to understanding complex information processing architectures that support flexible, context-appropriate behavior across diverse task domains.

2.1.4 Developmental Biology: Goal-Directed Adaptive Behavior

Developmental biologists see the slime mold as exhibiting clear goal-directed behavior through distributed cellular decision-making [Levin, 2023, Pezzulo and Levin, 2016]. Their framework emphasizes cellular competency, where individual protoplasmic regions function as "competent sub-agents" capable of sensing local nutrient gradients, responding adaptively to environmental conditions, coordinating with neighboring regions through chemical signaling, and maintaining homeostatic goals across multiple scales [Manicka and Levin, 2022, Pio-Lopez et al., 2023].

The system exhibits multi-scale organization where goal-directedness emerges at multiple levels. System-level goal achievement G_{system} can be understood as a function of local cellular goals g_i and coordination mechanisms C:

$$G_{system} = f(g_1, g_2, ..., g_n, C) \tag{2}$$

This framework predicts adaptive responses to novel perturbations beyond simple feedback, context-dependent behavior reflecting cellular "decisions," robustness through distributed intelligence rather than centralized control, and capacity for morphological computation across different problem types [Levin, 2012, Sultan et al., 2021].

The developmental biology compression treats goal-directedness as emergent from self-organizing networks, requiring neither explicit representation nor centralized planning [Steinberg, 1998, Lander, 2011]. This reflects the field's focus on understanding how complex organizational patterns arise through coordinated activities of simpler components, each maintaining local goals that collectively generate adaptive system-level behaviors.

2.1.5 Computer Science: Emergent Optimization Without Agency

Computer scientists analyze the system as implementing a distributed optimization algorithm through physical processes [Masterman et al., 2024]. The algorithmic analysis treats the slime mold as approximating parallel exploration through breadth-first search, gradient descent on path length through tube thickness dynamics, greedy optimization with implicit backtracking, and approximate solutions to Steiner tree problems. The computational complexity can be characterized, with solution time scaling as $T_{solution} \sim O(n^2 \log n)$ where n is the number of nodes in the network.

This framework predicts that solution quality depends on graph topology and initial conditions, performance degrades predictably with problem complexity, scaling behavior follows algorithmic complexity bounds, and the system exhibits equivalence to specific classes of optimization algorithms [Bryson, 2000, Vernon et al., 2007]. The computer science compression treats the system as instantiating optimization procedures without requiring "true" agency—it represents distributed processing that happens to solve problems efficiently.

2.1.6 Synthesis: Different Mathematics, Different Predictions

These four approaches generate different testable predictions. The control theory framework accurately predicts behavior in stable environments with familiar perturbations, with performance depending only on feedback parameters and no learning across trials [Powers, 1978, Mulder et al., 2018]. The cognitive science

framework predicts no genuine problem-solving, only reactive responses to local gradients with no memory of previous configurations [Cooper and Peebles, 2015]. The developmental biology framework better captures adaptive responses to novel challenges through context-dependent cellular decisions [Levin, 2023, Manicka and Levin, 2022]. The computer science analysis correctly bounds performance on structured problems according to algorithmic complexity [Masterman et al., 2024].

Empirical testing reveals that different predictions hold in different contexts. None of these frameworks is simply "wrong"—each captures different aspects of the system's behavior that prove relevant for different purposes. This exemplifies the central insight: agency modeling reflects compression strategies optimized for specific predictive goals, not discovery of inherent properties.

2.2 Bacterial Antibiotic Resistance: Strategic Behavior from Minimal Agents

2.2.1 The Phenomenon

Bacterial populations developing antibiotic resistance exhibit coordinated behaviors that superficially resemble strategic planning. Subpopulations differentiate into distinct functional roles: some cells express costly resistance mechanisms and share these with neighbors through horizontal gene transfer, others remain sensitive but benefit from the resistant population's activities, still others enter dormant states that survive antibiotic exposure through persistence rather than resistance. The population exhibits bet-hedging strategies, maintaining phenotypic diversity that ensures survival under uncertain environmental conditions. Colony organization adapts to antibiotic gradients, with resistant cells often positioning themselves at boundaries where they protect sensitive interior populations.

The question is whether this constitutes strategic agency. The behaviors resemble sophisticated military tactics—coordinated defense, sacrifice for group benefit, resource sharing, adaptive positioning. Yet no individual bacterium possesses knowledge of the larger strategy or explicit representation of group goals.

2.2.2 Evolutionary Game Theory: Population-Level Dynamics from Simple Rules

Evolutionary biologists model bacterial resistance through replicator dynamics and evolutionary game theory [Hofbauer and Sigmund, 1998, Broom and Cannings, 2010]. Individual bacteria are treated as agents employing simple reactive strategies—payoff-driven imitation of successful neighbors, myopic updating based on immediate fitness consequences, and frequency-dependent learning without explicit foresight. The mathematical framework uses replicator equations describing how strategy frequencies change proportional to their fitness advantage relative to the population average:

$$\frac{dx_i}{dt} = x_i(f_i(x) - \bar{f}(x)) \tag{3}$$

where x_i represents the frequency of strategy i, $f_i(x)$ represents the fitness of strategy i given the current population state x, and $\bar{f}(x)$ represents mean population fitness.

This framework predicts convergence to evolutionarily stable strategies (ESS) where no mutant strategy can invade, emergence of sophisticated population-level behaviors from minimal individual cognition, stable polymorphisms maintaining strategic diversity through frequency-dependent selection, and rapid adaptation through selection on standing genetic variation [Aoki and Feldman, 2014, Nowak et al., 2010]. The model requires no strategic reasoning, explicit memory beyond simple learning rules, or theory of mind. Sophisticated collective outcomes emerge through population dynamics operating on simple behavioral variants [Henrich and Boyd, 2002].

The evolutionary compression emphasizes sufficiency of minimal agency for generating adaptive population behaviors. Selection pressures act as an external computational system, obviating the need for complex internal cognition [Adami et al., 2016]. This reflects the field's recognition that evolutionary processes can generate apparently purposeful designs without requiring designers to possess foresight or explicit goals.

2.2.3 Behavioral Economics: Bounded Strategic Reasoning

Behavioral economists studying antibiotic resistance might instead apply level-k thinking models and cognitive hierarchy frameworks [Camerer et al., 2004, Crawford et al., 2013]. While bacteria obviously lack

conscious deliberation, the framework could model population-level strategic patterns through limited recursive reasoning. Level-0 bacteria might act non-strategically, expressing resistance randomly. Level-1 bacteria best-respond to level-0 patterns, expressing resistance when neighbors provide insufficient protection. Higher-level reasoning would involve more sophisticated anticipation of others' responses.

The mathematical framework would model belief distributions over thinking levels and predict behavior through bounded strategic reasoning:

$$P(\text{strategy } s|\text{level } k) = \text{BR}_k(P(\text{opponent strategies}|\text{levels } < k)) \tag{4}$$

where BR_k represents the best response given beliefs about lower-level opponents. This approach would predict systematic deviations from Nash equilibrium, limited depth of strategic sophistication, and heterogeneity in strategic thinking across the population [Alaoui and Penta, 2015, Wright and Leyton-Brown, 2017].

However, applying this framework to bacteria requires careful justification. The level-k framework was developed for bounded human cognition [Camerer et al., 2003], not microbial populations. Its value would lie in capturing strategic patterns without claiming bacteria possess conscious reasoning, treating strategic depth as an emergent population property rather than individual cognitive capability.

2.2.4 Cognitive Science: Rejecting Strategic Attribution

Cognitive scientists would reject strategic attribution to bacterial populations [Cooper and Peebles, 2015, Kriegeskorte and Douglas, 2018]. The framework emphasizes absence of necessary cognitive features: no mental representation of strategies or opponents, no working memory for tracking interaction history, no explicit goal structures beyond biochemical imperatives, no theory of mind or capacity to model others' states, and no symbolic processing or abstract reasoning. Observed behaviors result from biochemical feedback loops and gene regulatory networks, not strategic cognition [Doumas and Hummel, 2012, Fitch, 2014].

This perspective predicts that interventions targeting strategic thinking would fail—bacteria respond to immediate biochemical signals, not strategic anticipation. Population patterns reflect evolutionary selection on genetic variants, not learning from strategic experience. Apparent coordination emerges from shared genetic programs and chemical signaling, not explicit communication of strategic intent.

The cognitive science framework maintains strict criteria for attributing strategic reasoning, requiring explicit representational structures and goal-directed planning absent in bacterial systems. This reflects commitment to distinguishing genuine cognition from superficially similar behaviors arising through different mechanisms.

2.2.5 Synthesis: Strategic Patterns Without Strategic Agents

The bacterial resistance case reveals how strategic-looking patterns can emerge through multiple distinct mechanisms. The evolutionary framework demonstrates sufficiency of minimal agency—simple reactive rules plus selection pressure generate sophisticated population dynamics [Hofbauer and Sigmund, 1998, Aoki and Feldman, 2014]. The behavioral economics framework shows how strategic patterns could be captured through bounded reasoning models, though applying this to bacteria requires careful interpretation [Camerer et al., 2004]. The cognitive science framework rejects strategic attribution entirely, maintaining that genuine strategy requires cognitive structures bacteria lack [Cooper and Peebles, 2015].

These different compressions reflect different explanatory goals. Evolutionary biology seeks to understand population-level dynamics and adaptation [Broom and Cannings, 2010]. Behavioral economics focuses on strategic patterns and deviations from rationality [Crawford et al., 2013]. Cognitive science aims to identify genuine information-processing architectures [Kriegeskorte and Douglas, 2018]. Each framework proves useful for different purposes—none captures the complete truth, each reveals different aspects of the phenomenon.

2.3 Large Language Models: Agency in Novel Territory

2.3.1 The Phenomenon

Large language models like GPT-4 exhibit behaviors that challenge existing agency frameworks [Park et al., 2023]. They engage in complex dialogue maintaining conversational context, demonstrate apparent reasoning

through chain-of-thought processes, adapt their responses based on interaction history, exhibit consistency in following instructions while remaining flexible to novel requests, and generate sophisticated outputs across diverse domains. The question of whether these systems constitute agents, and which agency features they possess, remains contentious.

2.3.2 Computer Science: Architectural Analysis

Computer scientists analyze LLMs through architectural features and computational mechanisms [Masterman et al., 2024]. Recent surveys distinguish between purely reactive systems that process inputs independently, systems with memory maintaining context through attention mechanisms, and systems with strategic reasoning implementing planning through multi-step inference. The architectural analysis examines whether LLMs possess working memory through context windows, long-term memory through parameter storage, strategic reasoning through chain-of-thought prompting, goal-directedness through instruction following and reward modeling, and learning through fine-tuning and in-context learning [Mali, 2002, Vernon et al., 2007].

The framework predicts that LLM capabilities scale with architectural features—context window size limits working memory capacity, parameter count constrains knowledge storage, and inference computational budget determines reasoning depth. Performance depends on training data distribution, fine-tuning objectives, and prompt engineering. The analysis treats agency as an emergent property of architectural choices rather than a fundamental system feature [Bryson, 2000].

2.3.3 Cognitive Science: The Representation Question

Cognitive scientists debate whether LLMs possess genuine cognitive features or merely simulate them through statistical patterns [Kriegeskorte and Douglas, 2018, Cooper and Peebles, 2015]. The central question concerns mental representation—do LLMs maintain internal models of entities and relationships, or do they operate purely through pattern completion? The framework examines whether systems exhibit compositional structure in their representations, systematic generalization to novel combinations, explicit reasoning traceable through intermediate steps, genuine memory distinct from statistical associations, and theory of mind through modeling user beliefs and intentions [Doumas and Hummel, 2012, Fitch, 2014].

Current evidence suggests mixed results. LLMs demonstrate some compositional abilities but fail on systematic generalization tests. They exhibit reasoning-like behaviors but lack transparent internal processes. They maintain context but operate through attention mechanisms rather than explicit memory stores. The cognitive science framework remains agnostic, treating LLMs as systems whose cognitive status requires further empirical investigation [Sowa, 2011].

2.3.4 Behavioral Economics: Bounded Rationality in Machines

Behavioral economists examining LLM decision-making might apply bounded rationality frameworks [Camerer et al., 2003, Crawford et al., 2013]. Do LLMs exhibit systematic deviations from optimality similar to human cognitive constraints? Do they display heuristics and biases indicating computational limitations? The framework would analyze whether LLMs show anchoring effects, framing effects, limited strategic depth in game-theoretic scenarios, and satisficing rather than optimizing behavior [Simon, 1982].

Empirical studies suggest LLMs do exhibit bounded rationality patterns, though whether these reflect genuine computational constraints or merely replicate training data biases remains unclear. The bounded rationality framework proves useful for characterizing LLM behavior in decision-making contexts, regardless of underlying mechanisms [Kuzmanovic, 2021].

2.3.5 Synthesis: Boundary Cases Reveal Framework Limits

The LLM case exposes limitations in existing agency frameworks developed for biological or engineered systems. Computer science frameworks focus on architectural features without determining whether these constitute genuine agency [Masterman et al., 2024]. Cognitive science frameworks struggle to apply criteria developed for biological cognition to fundamentally different computational substrates [Kriegeskorte and Douglas, 2018]. Behavioral economics frameworks characterize decision patterns without resolving questions about underlying representations [Crawford et al., 2013].

These boundary cases motivate our systematic cross-domain analysis. By examining how established fields have successfully navigated agency modeling challenges in their respective domains, we can identify principles that might guide agency attribution in novel systems like LLMs. The systematic patterns we document in Section 3 reveal when different agency features become necessary versus sufficient, providing frameworks for evaluating new cases where existing approaches prove incomplete.

3 A Landscape of Agency Modeling Stances Across Domains

To build a more comprehensive understanding, we now explore how different scientific domains adopt distinct agency modeling stances, driven by their unique predictive goals. For each domain, we consider its typical compression strategy and the "agent features" it emphasizes, informed by syntheses of their respective literatures. Table 1 presents an illustrative mapping.

Table 1: Illustrative Postulates on the Necessity (N) or Sufficiency (S) of Model Features for Predictive Tasks in Diverse Domains, Synthesized from Literature Reviews

Model Feature (often from GT Stance)	**Behavioral Econ. & Experim. GT**	**Evolutionary Systems**	**Developmental Systems (Bio/Cog)**	**AI & Robotics (incl. Swarm)**	**Minimal Agency & Control Systems**	**Cognitive Science (General)**
Primary Predictive Goal(s)	Human choices in strategic settings.	Stability & spread of traits/strategies.	Ontogenetic pathways, emergence of complexity.	Goal achievement in dynamic environments.	System stability, setpoint tracking.	Mechanisms of thought & behavior.
Typical Compression Strategy	Humans as boundedly rational, heuristic-driven.	Agents as replicators/ learners under selection.	Systems as self-organizing, goal-directed networks.	Designed goal-achievers, problem-solvers, learners.	Goal-seeking via feedback loops.	Mind as information processor.
Strategic Reasoning	N (limited/ heuristic-based)	S (simple reactive rules often suffice)	S (emergent goal-direction, not necessarily strategic)	N (planning, MAS coordination)	S (simple error-correction logic)	N (problem-solving, decision-making)
Memory	N (learning, context)	S (minimal for simple learning/imitation)	N (state tracking, learning)	N (for POMDPs, learning from experience)	S (minimal state for feedback)	N (working, LTM, episodic are core)
Theory of Mind	S (level-k, cognitive hierarchy models)	- (rarely explicit; often emergent social learning)	S (for advanced social cognitive development, not basic morphogenesis)	S (increasingly for MAS, HRI; not universal)	- (not applicable)	N (core for social cognition)
Self-Reflection	-	-	-	S (in advanced learning agents, e.g., Reflexion)	-	S (metacognition)
Group Cognition / Collective Behavior	S (social preferences, norms)	N (multi-level selection, cultural transmission)	S (cell-cell coordination, emergent tissue properties)	N (swarm intelligence, MAS coordination)	-	S (social norms, distributed cognition)
Comp. Boundedness	N	N (implicit in rule simplicity)	N (biological processing limits)	N (practical AI design constraint)	N (physical/ computational limits)	N (central to many cognitive architectures)
Goal-Directedness	N	N (fitness)	N	N	N	N
Learning/ Adaptation	N	N	N	N	S	N
Feedback Control	-	S (environmental feedback drives selection)	N (homeostatic loops)	N (RL, adaptive control)	N	N (error-driven learning)
Symbolic Processing	S (in some cognitive models)	-	S (developmental "programs")	S (hybrid AI architectures)	-	S (classical cognitive architectures)

3.1 Behavioral Economics & Experimental Game Theory

A Behavioral Economist might say: "Our primary goal is to predict and explain how real humans make decisions in strategic, often economic, contexts. We start with the predictions of classical game theory, which assumes full rationality, but find that human behavior systematically deviates. Therefore, our modeling stance compresses human agency by incorporating **Bounded Rationality** and specific **Cognitive Constraints**; these are necessary features of our models [**?**Crawford et al., 2013]. We observe that people use **Heuristics** and simplified decision strategies rather than performing perfect optimization. While some **Strategic Reasoning** is evident, it's often limited to a few levels of recursion, as captured by level-k or cognitive hierarchy models [**?**]. **Memory** and **Learning** are crucial in repeated interactions, but again, often bounded. Explicit, deep **Theory of Mind** isn't always necessary to model behavior, as simpler rules or fairness considerations can drive choices."

3.1.1 Evolution of Modeling Approaches

Behavioral economics emerged from systematic empirical testing of classical game theory predictions in controlled laboratory settings. Early experimental work in the 1980s and 1990s consistently documented deviations from Nash equilibrium predictions across diverse strategic contexts [Camerer, 1997]. This empirical evidence led to a gradual shift in modeling approaches, moving from assumptions of full rationality toward incorporating cognitive constraints.

The field has converged around several key modeling frameworks. **Bounded rationality** refers to decision-making that is rational within cognitive limitations, rather than globally optimal. **Level-k thinking** models assume players engage in limited recursive reasoning, where level-0 players act non-strategically, level-1 players best-respond to level-0 players, and so forth. **Cognitive hierarchy models** extend this by assuming players have beliefs about the distribution of thinking levels in the population [Camerer et al., 2004].

3.1.2 Goal-Directedness and Analysis Approaches

The goal-directedness in behavioral economics modeling is primarily **predictive** rather than mechanistic. Researchers seek to predict aggregate behavior patterns in strategic contexts without necessarily explaining the underlying cognitive mechanisms. This leads to a predominantly **black box** approach where internal cognitive processes are compressed into simplified behavioral rules or probability distributions over actions.

However, some recent work incorporates **white box** elements by connecting behavioral patterns to specific cognitive mechanisms. For instance, computational neuroscience approaches attempt to link level-k thinking patterns to measurable neural activity [Griessinger et al., 2018]. The choice between black box and white box analysis depends on the specific research question and available measurement techniques.

3.1.3 Strategic Reasoning: Necessary vs. Important

Strategic Reasoning serves different roles depending on the modeling context. In single-shot games, some form of strategic reasoning is **necessary** for the model to capture interdependent decision-making. However, the depth of strategic reasoning varies significantly across experimental contexts.

Meta-analytic evidence suggests most experimental participants operate at level-1 or level-2 thinking, with substantial heterogeneity [**?**]. This bounded strategic reasoning is **important** for capturing realistic behavior patterns, but unlimited recursive reasoning is neither necessary nor empirically supported. The field has thus settled on models that incorporate limited strategic sophistication as both necessary and sufficient for most predictive purposes.

3.1.4 Memory and Learning Mechanisms

Memory becomes necessary in repeated strategic interactions where history affects optimal strategies. However, behavioral economics typically employs simplified memory models rather than detailed cognitive architectures. Common approaches include:

- Experience-weighted attraction learning, where players weight past outcomes by recency and similarity

- Belief learning models that update probabilistic beliefs about opponents' strategies
- Reinforcement learning with limited memory windows

These models treat memory as a **sufficient** approximation for capturing adaptive behavior without modeling detailed memory mechanisms. The distinction between necessary and important becomes clear here: some memory mechanism is necessary for repeated games, but detailed memory architecture is not important for most behavioral predictions.

3.1.5 Theory of Mind: Context-Dependent Deployment

Theory of Mind in behavioral economics refers to players' ability to model others' mental states and reasoning processes. However, the field has found that explicit theory of mind is often not necessary for explaining strategic behavior. Many experimental findings can be explained through simpler mechanisms:

- Social preference models that incorporate fairness or reciprocity concerns
- Imitation or social learning rules
- Focal point reasoning in coordination games

Theory of mind becomes more important in complex games with incomplete information or repeated interactions where reputation matters. The level-k thinking framework provides a **sufficient** approximation of theory of mind for most experimental contexts without requiring detailed modeling of recursive mental state attribution.

3.1.6 Computational Boundedness as Design Principle

Computational Boundedness is both necessary and central to behavioral economics modeling. Unlike other domains where computational limits are treated as constraints to work around, behavioral economics treats them as fundamental features that explain observed behavior patterns.

The field has systematically documented how computational constraints shape strategic behavior:

- Limited working memory affects the depth of strategic reasoning
- Time pressure reduces strategic sophistication
- Cognitive load interferes with theory of mind deployment
- Complexity of the strategic environment determines which heuristics are employed

This makes computational boundedness a **necessary** feature for realistic models, distinguishing behavioral economics from classical game theory approaches that assume unlimited computational capacity.

3.1.7 Experimental Paradigms and Agent-Environment Boundaries

The agent-environment boundary in behavioral economics is defined by experimental design. Researchers manipulate information availability, payoff structures, and communication opportunities to isolate specific aspects of strategic reasoning. This experimental approach allows systematic testing of when different agency features become necessary or sufficient.

Common experimental paradigms include:

- One-shot vs. repeated games to test memory and learning requirements
- Varying information structures to examine theory of mind deployment
- Cognitive load manipulations to study computational boundedness effects
- Cross-validation across different game types to test model generalizability

The controlled nature of experiments enables researchers to establish when specific agency features are necessary for adequate prediction versus when simpler models prove sufficient.

3.2 Evolutionary Systems (Biology, Evolutionary Game Theory, Cultural Evolution)

An Evolutionary Theorist might state: "We aim to predict the long-term dynamics of traits and strategies in populations under selection pressure. Our modeling stance views entities—genes, individuals, groups, or even cultural traits—as units subject to replication and selection, where 'fitness' or a similar utility is the currency. Often, **Minimal Agency** suffices for our models [Hofbauer and Sigmund, 1998]. Agents are frequently modeled with reactive strategies (e.g., payoff-driven imitation, myopic updates) and basic **Learning Rules**, which are sufficient to explain the emergence of Evolutionary Stable Strategies (ESS) and population-level outcomes via mechanisms like replicator dynamics [**?**]. While **Memory** can be incorporated for repeated interactions, complex cognitive features like explicit **Theory of Mind** or deep strategic reasoning are often abstracted away, as selection can favor simpler, robust heuristics. Cultural evolution models might incorporate more complex cognitive features like bounded rationality or inferential processes for phenomena like cumulative adaptation [Henrich and Boyd, 2002]. Multi-level selection theories explore how agency and goal-directedness can be attributed across different levels of biological organization [Broom and Cannings, 2010], where group-level fitness can drive the evolution of individual traits."

This characterization reveals one of the most intellectually fascinating aspects of evolutionary thinking: the elegant parsimony with which complex behavioral phenomena can emerge from strikingly simple underlying mechanisms. The evolutionary perspective offers a unique window into agency modeling because it must simultaneously account for both the constraints of biological reality and the mathematical necessity of population-level dynamics.

3.2.1 The Architecture of Minimal Agency

The concept of minimal agency in evolutionary systems represents perhaps the most theoretically profound compression strategy across all scientific domains. When we examine the theoretical foundations laid out by researchers like Sandholm [**?**] and Hofbauer and Sigmund [Hofbauer and Sigmund, 1998], we discover that the overwhelming majority of evolutionary outcomes can be predicted using agents that exhibit only the most basic reactive capabilities.

This minimal agency typically consists of three core elements: **payoff sensitivity** (the ability to distinguish between more and less successful outcomes), **imitation capacity** (the ability to copy successful strategies from others), and **myopic updating** (the tendency to adjust behavior based on immediate rather than long-term consequences). The mathematical elegance of this approach lies in how these simple individual-level mechanisms, when embedded within population dynamics governed by replicator equations, generate remarkably sophisticated collective behaviors.

The theoretical reviews consistently demonstrate that this minimal agency framework proves **sufficient** for modeling evolutionary stable strategies across diverse biological contexts [Aoki and Feldman, 2014, Nowak et al., 2010]. The key insight here is that evolutionary pressures themselves serve as a kind of external computational system, obviating the need for complex internal cognition. When selection operates efficiently, simple reactive strategies can outcompete more sophisticated cognitive approaches because they avoid the metabolic and developmental costs associated with complex neural architectures.

3.2.2 Memory and Learning Rules: Strategic Minimalism

The treatment of **Memory** and **Learning** in evolutionary models reveals a fascinating tension between computational efficiency and adaptive capability. Unlike other domains where memory systems are modeled as general-purpose storage and retrieval mechanisms, evolutionary approaches treat memory as a specialized adaptation that emerges only when the benefits outweigh the costs.

Most foundational evolutionary models operate without explicit memory mechanisms, relying instead on what might be termed "environmental memory"—where the current state of the population itself encodes relevant historical information about strategy success [Broom and Cannings, 2010]. This approach treats memory as **unnecessary** for basic evolutionary dynamics, since frequency-dependent selection naturally captures the cumulative outcomes of past strategic interactions.

However, the field has identified specific contexts where memory becomes **necessary** rather than merely beneficial. In temporally variable environments, as analyzed by Aoki and Feldman [Aoki and Feldman,

2014], agents require some capacity to track environmental changes and adjust their behavioral strategies accordingly. The critical insight is that even this "memory" is typically modeled as simple state-dependent switching rules rather than complex episodic storage systems.

Learning rules in evolutionary contexts exhibit a similar minimalist elegance. Rather than modeling learning as conscious information processing, evolutionary approaches treat it as genetically or culturally programmed response patterns that can themselves evolve. This represents a profound reframing: learning capacity becomes a trait subject to selection rather than a cognitive process requiring detailed mechanistic modeling.

3.2.3 The Cultural Evolution Threshold

One of the most intellectually compelling findings from the theoretical literature concerns the emergence of a clear complexity threshold when evolutionary models incorporate cultural transmission mechanisms. While genetic evolution models consistently demonstrate the sufficiency of minimal agency, cultural evolution frameworks reveal fundamentally different computational requirements.

Henrich and Boyd's seminal theoretical analysis [**?**] demonstrates that cultural transmission creates selection pressures for cognitive capabilities that are unnecessary in purely genetic evolution. Their models suggest that **inferential processes** and **cognitive attractors** become necessary features when modeling cultural evolution, particularly for phenomena like cumulative cultural adaptation.

This threshold emerges from the fundamental differences between genetic and cultural inheritance systems. Genetic transmission operates through high-fidelity molecular copying mechanisms that require minimal cognitive input from the organism. Cultural transmission, by contrast, relies on social learning processes that inherently involve interpretation, inference, and active reconstruction of information. This creates selection pressures for cognitive systems capable of extracting underlying patterns from noisy social input—a computational challenge that pushes beyond the minimal agency framework.

The implications for agency modeling are profound. Cultural evolution models require agents capable of **bounded rationality**, where individuals can make reasonable inferences under uncertainty while remaining computationally tractable. This represents a qualitative shift from reactive strategies to what might be termed "constructive strategies"—approaches that actively generate novel behavioral variants through cognitive processing rather than merely selecting among existing alternatives.

3.2.4 Multi-Level Selection and Emergent Agency

Perhaps the most theoretically ambitious frontier in evolutionary agency modeling concerns multi-level selection theory and the emergence of agency across different scales of biological organization. The theoretical frameworks developed by researchers like Van Cleve [**?**] and Broom and Cannings [Broom and Cannings, 2010] reveal how agency properties can emerge at group levels through selection processes operating simultaneously across individual and collective scales.

This multi-level perspective fundamentally challenges traditional assumptions about where agency "resides" in biological systems. Rather than treating individual organisms as the natural unit of agency, multi-level selection theory suggests that goal-directedness and strategic behavior can emerge at any level where there is sufficient variation, inheritance, and differential fitness.

The mathematical frameworks required for multi-level selection reveal interesting patterns in agency feature requirements. **Group Cognition** emerges as a necessary feature when modeling collective behaviors that cannot be reduced to individual-level strategies. Examples include coordinated collective movements, group decision-making processes, and the evolution of communication systems that benefit group-level fitness at potential costs to individual fitness.

These models demonstrate how **Strategic Reasoning** can emerge at collective levels even when individual agents exhibit only minimal cognitive capabilities. The classic example involves social insect colonies, where sophisticated collective strategies emerge from interactions among individuals following simple behavioral rules. This suggests that strategic sophistication can be distributed across multiple agents rather than requiring complex individual cognition.

3.2.5 Replicator Dynamics and Strategy Representation

The mathematical formalism of replicator dynamics provides perhaps the clearest window into how evolutionary approaches compress agency into tractable mathematical representations. The elegance of the replicator equation—where strategy frequencies change proportional to their fitness advantage relative to the population average—reveals how complex strategic interactions can be captured without modeling detailed cognitive processes.

This mathematical framework treats strategies as abstract beha

3.3 Developmental Systems (Biology & Cognitive Development)

A Developmental Systems Biologist/Psychologist might explain: "Our goal is to understand and predict how complex organisms and cognitive capabilities arise from simpler beginnings through ontogenetic processes. We model systems as active, self-organizing networks where **Goal-Directedness** is often an emergent property rather than a pre-programmed one. For instance, in morphogenesis, cells and tissues exhibit behaviors that can be described using pragmatic teleological language—referring to developmental programs, targets, and setpoints—which are useful compressions of complex underlying dynamics [**?**Lander, 2011]. These systems often rely on **Feedback Control** and homeostatic loops across multiple scales, from cellular to organismal. While not 'strategic reasoning' in the adult human sense, cellular decision-making and interactions are crucial. In cognitive development, we see a sequential acquisition of capabilities: basic **Learning Mechanisms** and **Memory** are necessary for the later emergence of a sophisticated **Theory of Mind** [**?**]. The agent-environment boundary is co-constructed and changes throughout development, with the organism actively shaping and being shaped by its environment."

This characterization captures something profoundly fascinating about developmental systems: they represent perhaps the most elegant demonstration of how sophisticated, apparently purposeful behaviors can emerge from the coordinated interactions of components that individually possess no knowledge of the larger design. The intellectual journey of developmental biology over the past several decades reveals a field grappling with one of the most fundamental questions in science—how does complex organization arise from simple beginnings?

3.3.1 The Great Paradigm Shift: From Genetic Programs to Emergent Systems

The theoretical landscape of developmental biology has undergone a remarkable transformation that mirrors broader shifts across the life sciences. The classical view, dominated by genetic determinism, conceived of development as the unfolding of pre-programmed instructions encoded in DNA. This framework treated genes as master controllers orchestrating development through hierarchical command structures, with cellular behavior reduced to the passive execution of genetic directives.

Contemporary developmental systems theory has fundamentally challenged this reductionist paradigm. The accumulated evidence from decades of experimental work has revealed that developmental outcomes emerge from dynamic interactions between multiple levels of biological organization, none of which can be understood in isolation [Sultan et al., 2021]. This shift represents more than a mere refinement of existing models—it constitutes a fundamental reconceptualization of how biological agency operates during development.

The new paradigm recognizes that **Goal-Directedness** in developmental systems represents an emergent property that arises from the self-organizing dynamics of complex networks rather than from pre-specified genetic instructions. This insight has profound implications for how we model agency in biological systems, suggesting that purposeful behavior can emerge without centralized control or explicit programming.

3.3.2 Multi-Scale Homeostatic Architecture

One of the most intellectually compelling discoveries in contemporary developmental biology concerns the hierarchical organization of homeostatic control systems that operate across multiple scales of biological organization. These homeostatic loops represent a form of distributed agency where goal-directedness emerges through the coordinated operation of feedback mechanisms spanning molecular, cellular, tissue, and organismal levels.

The TAME framework (Transforming Agent-based Models to Explain homeostasis) developed by Pio-López and colleagues [**??**] provides a particularly elegant mathematical formalization of how cellular-level homeostatic goals can scale to produce anatomical-level coordinated behaviors. Their agent-based cellular automata models demonstrate how individual cells, each maintaining local homeostatic setpoints, can collectively generate tissue-level patterns that appear purposefully directed toward specific morphological outcomes.

This multi-scale homeostatic architecture reveals that **Memory** in developmental systems operates very differently than in cognitive or behavioral contexts. Rather than serving as a repository for past experiences, developmental memory functions as a dynamic maintenance system that preserves essential organizational relationships across scales. Cellular memory systems maintain epigenetic states that encode positional information and developmental history, while tissue-level memory emerges from the stable maintenance of morphological patterns through homeostatic feedback loops.

The scaling properties of these homeostatic systems demonstrate that **Feedback Control** becomes both necessary and sufficient for explaining coordinated developmental outcomes. Individual cells need not possess knowledge of global developmental goals—their local homeostatic activities, when properly coordinated through stress propagation and information flow, spontaneously generate the appropriate tissue-level responses.

3.3.3 Cellular Competency and Distributed Decision-Making

Perhaps the most revolutionary insight emerging from contemporary developmental biology concerns the recognition that individual cells exhibit sophisticated forms of decision-making that contribute to collective developmental outcomes. This cellular competency represents a form of minimal agency that operates through information processing, environmental sensing, and adaptive response mechanisms.

Levin's framework of multiscale competency architecture [**?**] conceptualizes development as emerging from the coordinated activities of "competent sub-agents" operating at multiple organizational levels. Individual cells function as information-processing entities capable of interpreting environmental cues, maintaining internal goals (homeostatic setpoints), and adjusting their behavior in response to changing conditions. This cellular agency, while not involving conscious deliberation, nevertheless exhibits many of the functional characteristics associated with goal-directed behavior.

The theoretical frameworks developed by Manicka and Levin [Manicka and Levin, 2022] demonstrate how **Strategic Reasoning** can emerge at the cellular level through causal network architectures that enable cells to predict and respond to developmental contingencies. These cellular "decisions" involve weighing multiple environmental inputs, integrating historical information (through epigenetic mechanisms), and selecting among alternative developmental pathways based on context-dependent criteria.

This distributed decision-making architecture reveals that **Group Cognition** emerges naturally in developmental systems through the coordination of cellular competencies. Tissues and organs exhibit collective intelligence properties that arise from the integration of individual cellular responses, enabling adaptive responses to environmental perturbations that maintain developmental robustness while allowing for morphological plasticity.

3.3.4 The Pragmatic Justification of Teleological Language

One of the most intellectually sophisticated aspects of contemporary developmental biology concerns its thoughtful approach to teleological language. Unlike fields that either embrace or reject purposive descriptions wholesale, developmental biologists have developed nuanced justifications for when and why goal-directed language proves scientifically useful.

The pragmatic use of teleological terminology in developmental biology reflects a deep understanding that purposive language can capture real features of biological systems without implying conscious intent or supernatural design [**?**Steinberg, 1998]. When developmental biologists refer to cellular "goals," tissue "targets," or developmental "programs," they are employing these terms as compressed descriptions of observable regulatory behaviors that maintain specific organizational states despite environmental perturbations.

This approach treats teleological language as an empirical description of system properties rather than a metaphysical commitment. Cells exhibit goal-directed behavior in the functional sense that they actively maintain specific physiological states, respond adaptively to perturbations, and coordinate their activities to

achieve collective outcomes. The teleological description captures these regulatory dynamics more efficiently than purely mechanistic accounts while remaining grounded in observable phenomena.

The sophisticated use of purposive language in developmental biology demonstrates how scientific fields can employ agent-like descriptions without anthropomorphizing their subject matter. The key insight is that goal-directedness represents a legitimate level of biological organization that emerges from underlying mechanistic processes while exhibiting autonomous causal powers that cannot be reduced to lower-level descriptions.

3.3.5 Computational Frameworks for Developmental Agency

The mathematical and computational approaches employed in developmental biology reveal fascinating insights into how different modeling frameworks capture different aspects of agency and self-organization. The field has converged around several key computational paradigms, each offering unique perspectives on how developmental agency emerges and operates.

The variational free energy framework, adapted from neuroscience by Friston and colleagues [Friston et al., 2015], treats developmental systems as active inference engines that minimize prediction error through both perceptual updating and action. This framework suggests that morphogenetic processes can be understood as forms of **Computational Boundedness** where biological systems optimize their organizational states within the constraints imposed by physical and chemical limitations.

Agent-based modeling approaches, exemplified by the cellular automata frameworks developed by Pio-López and colleagues, reveal how local interaction rules can generate global coordination patterns without centralized control. These models demonstrate that sophisticated collective behaviors can emerge from relatively simple individual behavioral rules, suggesting that complex developmental outcomes may result from the coordination of minimal cellular competencies rather than sophisticated individual planning.

The cybernetic frameworks increasingly employed in developmental biology treat biological systems as recursive control hierarchies where feedback loops operate across multiple temporal and spatial scales. This approach reveals how **Self-Reflection** can emerge at the tissue and organ levels through recursive feedback relationships that enable biological systems to monitor and adjust their own organizational states.

3.3.6 Information Flow and Cross-Scale Integration

The contemporary understanding of developmental agency increasingly emphasizes the crucial role of information flow in coordinating activities across different scales of biological organization. This information-theoretic perspective reveals how developmental systems maintain coherent global organization while preserving the autonomy of local regulatory processes.

Collinet and Lecuit's framework for understanding morphogenetic information flow [Collinet and Lecuit, 2021] demonstrates how developmental systems integrate both deterministic and self-organized mechanisms to achieve robust morphological outcomes. Their analysis reveals that successful development requires the coordination of multiple information processing streams, each operating according to different organizational principles but contributing to coherent collective outcomes.

This multi-stream information architecture suggests that **Theory of Mind** in developmental contexts operates through sophisticated intercellular communication systems that enable cells to model and predict the states of their neighbors. While this cellular "theory of mind" does not involve conscious mental state attribution, it does require sophisticated information processing capabilities that enable cells to infer the physiological states and likely behaviors of surrounding cells.

The cross-scale integration mechanisms identified in developmental biology demonstrate how local cellular competencies can be coordinated to generate tissue-level and organ-level behaviors that appear purposefully directed toward specific morphological goals. This integration occurs through stress propagation mechanisms, chemical signaling networks, and mechanical feedback systems that create coherent information flow across multiple scales of organization.

3.3.7 Temporal Dynamics and Developmental Memory

The temporal aspects of developmental agency reveal particularly sophisticated forms of biological information processing that challenge traditional distinctions between memory, learning, and inheritance. Devel-

opmental systems maintain multiple forms of temporal information that enable them to coordinate current activities with past developmental history and future morphological requirements.

Epigenetic memory systems function as cellular-level information storage mechanisms that maintain positional and temporal information across cell divisions. These systems enable individual cells to "remember" their developmental history and use this information to guide future developmental decisions. This biological memory operates very differently from cognitive memory systems—it maintains essential organizational information rather than episodic experiences, and it functions through molecular mechanisms rather than neural networks.

The temporal coordination of developmental processes demonstrates that **Learning** in developmental contexts operates through adaptive regulatory mechanisms that enable biological systems to adjust their organizational strategies based on environmental feedback. This developmental learning involves the modification of gene expression patterns, the adjustment of cellular behavioral rules, and the refinement of intercellular communication networks in response to developmental contingencies.

These temporal dynamics reveal that developmental agency emerges through the integration of inherited organizational information, current environmental conditions, and adaptive regulatory responses. The resulting systems exhibit sophisticated forms of biological intelligence that enable them to navigate complex developmental challenges while maintaining essential organizational relationships across multiple scales and time periods.

The profound implications of these discoveries extend far beyond developmental biology itself, offering insights into the fundamental nature of biological agency, the emergence of complex organization from simple components, and the relationship between mechanistic processes and purposeful behaviors in living systems.

3.4 Artificial Intelligence & Robotics (including Swarm/Collective AI)

An AI Researcher/Roboticist might articulate: "Our primary predictive (and design) goal is to create artificial systems that can achieve complex goals in dynamic, often partially observable, and sometimes multi-agent environments. The modeling stance varies greatly depending on the task. For complex planning or navigation in POMDPs, models explicitly incorporating **Strategic Reasoning** (e.g., search algorithms), internal **Memory** (like belief states in Bayesian RL), and robust perception are necessary features [Masterman et al., 2024]. In multi-agent systems (MAS), effective coordination or competition often requires agents to model others, making some form of **Theory of Mind** a highly beneficial, if not always strictly necessary, feature for sophisticated interaction. **Learning and Adaptation** are almost always necessary for robust performance in novel situations. **Computational Boundedness** is an unavoidable practical constraint that necessitates trade-offs in agent architectures (reactive, deliberative, hybrid) [**?**]. Swarm intelligence demonstrates how complex collective behavior (**Group Cognition**) can emerge from individuals with minimal cognitive features, relying on environmental mediation and simple interaction rules rather than explicit ToM."

The artificial intelligence and robotics literature reveals a field explicitly engaged in agency design decisions, where researchers must make deliberate choices about which cognitive capabilities to implement based on systematic analysis of task requirements and computational constraints. Unlike other domains examined in this paper, AI explicitly treats agency features as engineering parameters to be optimized rather than as natural phenomena to be observed.

3.4.1 Architectural Paradigms and Agency Feature Selection

Literature reviews in AI agent architecture reveal several distinct paradigmatic approaches, each emphasizing different combinations of agency features based on theoretical commitments and empirical performance [Wooldridge and Jennings, 1995, Vernon et al., 2007]. Table 2 summarizes the primary architectural approaches and their characteristic agency feature requirements.

The cognitivist paradigm, exemplified by Belief-Desire-Intention (BDI) architectures, treats **Strategic Reasoning** and explicit **Memory** as necessary features for agents operating in structured environments where symbolic planning provides computational advantages [Wooldridge and Jennings, 1995, **?**]. These systems implement explicit goal representation and intention manipulation, requiring sophisticated internal models of environmental states and action consequences.

Table 2: AI Agent Architecture Types and Associated Agency Feature Requirements

Architecture Type	**Core Features**	**Environmental Requirements**	**Key Agency Features**
Cognitivist	Symbolic reasoning, explicit goal representation, BDI models	Structured, well-defined environments	**Strategic Reasoning**, **Memory**, goal-directedness
Behavior-based	Reactive behaviors, environmental coupling, minimal internal state	Dynamic, unstructured environments	Adaptive behavior, sensorimotor coordination
Emergent systems	Embodiment, sensorimotor loops, self-organization	Rich, interactive environments	**Learning**, autonomous development, robustness
Hybrid systems	Integration of symbolic and emergent features	Flexible environments	**Strategic Reasoning**, **Learning**, coordination

Behavior-based robotics, by contrast, demonstrates that many tasks can be accomplished with minimal agency features [**?**]. These architectures rely primarily on reactive behavioral coupling between perception and action, treating complex internal representations as computationally unnecessary for many real-time robotics applications. The success of behavior-based approaches suggests that **Strategic Reasoning** and **Memory** are sufficient rather than necessary for many embodied intelligence tasks.

Emergent systems approaches emphasize **Learning** and adaptation as primary agency features, with complex behaviors arising from self-organizing dynamics rather than explicit planning [**??**]. These systems treat embodiment and sensorimotor coordination as fundamental, suggesting that sophisticated cognitive features can emerge from simpler substrate capabilities under appropriate environmental conditions.

3.4.2 Task-Environment Coupling and Computational Requirements

The AI literature demonstrates systematic relationships between environmental characteristics and agency feature requirements. Reviews consistently identify environmental observability as a primary determinant of architectural necessity [Masterman et al., 2024, Vernon et al., 2007].

In fully observable environments, agents can operate effectively with reactive strategies that couple perception directly to action. **Memory** becomes unnecessary when all relevant information for decision-making is available in the current environmental state. However, partial observability creates mathematical requirements for internal state maintenance, making **Memory** systems necessary for optimal performance in POMDP environments.

Multi-agent environments introduce additional computational challenges requiring coordination mechanisms. However, the literature reveals that explicit **Theory of Mind** modeling is rarely implemented in successful multi-agent systems [**??**]. Instead, coordination typically occurs through environmental mediation or direct communication protocols, suggesting that Theory of Mind is sufficient rather than necessary for most multi-agent coordination tasks.

3.4.3 Computational Boundedness and Architectural Trade-offs

Computational Boundedness emerges as a universal constraint in AI systems, fundamentally shaping architectural decisions across all paradigms. The literature documents systematic trade-offs between cognitive sophistication and computational efficiency [**?**Bryson, 2000].

Reactive architectures excel in scenarios requiring rapid responses within strict temporal deadlines but prove insufficient for tasks requiring long-term planning or complex goal coordination. Deliberative architectures can solve complex planning problems but may exceed computational budgets for real-time applications. Hybrid architectures attempt to capture benefits of both approaches through modular organization, treating computational complexity as a controllable parameter rather than a fixed constraint.

Recent advances in agent architecture design increasingly employ attention mechanisms and resource allocation strategies that dynamically adjust reasoning complexity based on task demands [Masterman et al., 2024]. These approaches treat agency features as computational resources to be allocated rather than fixed architectural commitments.

3.4.4 Swarm Intelligence and Minimal Individual Agency

Swarm intelligence research demonstrates that sophisticated **Group Cognition** can emerge from interactions between agents with minimal individual cognitive capabilities [**??**]. These systems achieve coordination through stigmergic mechanisms where agents modify shared environmental structures rather than maintaining explicit models of other agents' states.

The swarm intelligence literature reveals systematic relationships between individual agency complexity and collective capability. Simple agents employing basic sensorimotor coupling and environmental modification rules can generate complex collective behaviors including optimization, construction, and exploration tasks. This suggests that sophisticated collective intelligence does not require sophisticated individual cognition, challenging assumptions about the necessity of complex agency features for intelligent behavior.

Environmental mediation emerges as a primary coordination mechanism in swarm systems, where agents coordinate indirectly through environmental traces rather than direct communication or explicit mental state modeling. This approach sidesteps computational complexity associated with Theory of Mind while maintaining effective collective coordination.

3.4.5 Learning and Adaptation Requirements

Learning and Adaptation appear as necessary features across all AI architectural paradigms, though implementation approaches vary significantly [Vernon et al., 2007, **?**]. Reactive systems typically employ parameter optimization approaches within fixed behavioral structures, while deliberative systems require mechanisms for updating symbolic knowledge representations and modifying planning strategies.

The literature documents different learning requirements across individual and collective systems. Individual learning focuses on parameter adaptation and strategy modification within single agents, while collective learning involves coordination mechanisms that enable populations of agents to adapt their coordination strategies over time.

Hybrid architectures present particular challenges for learning system design, requiring coordination of adaptation mechanisms across architectural layers with different representational formats and temporal dynamics. Recent work explores meta-learning approaches that can adapt learning strategies themselves based on environmental demands.

3.4.6 Theory of Mind: Implementation Patterns and Alternatives

Despite theoretical expectations that multi-agent coordination would require explicit **Theory of Mind** capabilities, the AI literature reveals limited implementation of explicit mental state modeling [**?**]. Most successful multi-agent systems achieve coordination through alternative mechanisms that avoid the computational complexity of recursive mental state attribution.

Environmental mediation serves as the primary alternative to explicit Theory of Mind in multi-agent systems. Agents coordinate by modifying shared environmental structures, creating distributed information processing systems that enable sophisticated collective behaviors without explicit mental state modeling. This approach treats the environment as an external memory system that mediates coordination rather than requiring agents to maintain internal models of other agents' cognitive states.

Where Theory of Mind capabilities are implemented, they typically involve shallow, task-specific modeling rather than general-purpose mental state attribution. Competitive scenarios and coalition formation tasks show some benefit from explicit agent modeling, but even in these contexts, the modeling tends to focus on observable behavioral patterns rather than inferred mental states.

3.4.7 Synthesis: Agency as Engineering Parameter

The AI and robotics literature treats agency features as engineering parameters to be selected based on systematic analysis of task requirements, environmental constraints, and computational budgets. This engineering perspective provides explicit validation of the modeling stance framework proposed in this paper, demonstrating how different agency feature combinations prove necessary or sufficient depending on specific predictive and performance goals.

The field's systematic exploration of architectural trade-offs reveals general principles about relationships between environmental complexity, computational constraints, and agency feature requirements. These principles extend beyond robotics applications to any domain where intelligent behavior must be engineered rather than evolved, providing empirical validation for the theoretical framework developed in this paper.

Most significantly, the AI literature demonstrates that sophisticated intelligent behavior can emerge from careful organization of simple computational components, supporting the paper's central thesis that agency features represent modeling compressions rather than fundamental cognitive requirements. The success of minimal agency approaches in swarm intelligence and behavior-based robotics provides existence proofs that complex collective behaviors can arise without sophisticated individual cognitive architectures.

3.5 Minimal Agency & Control Systems

A Control Theorist or Cyberneticist might say: "Our focus is on understanding and designing systems that exhibit goal-directed behavior, typically maintaining stability or tracking reference signals despite disturbances. Our modeling stance compresses complex dynamics into feedback loops and error-correction mechanisms. The minimal criteria for what we might call 'agency' in these systems are the presence of **Feedback Control** (often negative feedback), a definable **Goal-State** (or reference value), and mechanisms for **Error Correction** and ensuring **System Stability** [**?**Powers, 1978]. Sophisticated **Strategic Reasoning** or **Theory of Mind** are not necessary for this foundational level of goal-seeking. Simple forms of **Memory** (e.g., the state in an integral controller) can be sufficient for certain control objectives. The agent-environment boundary is clearly defined by sensors (inputs) and actuators (outputs), and the system's purposiveness arises from its structural organization to reduce discrepancies from its goal state."

What emerges from the control theory and cybernetics literature is perhaps the most mathematically precise articulation of minimal agency found across any scientific domain. Here we encounter a field that has spent decades systematically exploring the absolute minimum requirements for goal-directed behavior, stripping away all unnecessary complexity to reveal the fundamental mathematical skeleton that underlies purposive action.

3.5.1 The Foundational Triad: Feedback, Reference, and Error

The theoretical literature converges with remarkable consistency around three fundamental components that constitute the irreducible core of goal-directed systems. Table 3 summarizes these essential features as identified across different control-theoretic frameworks.

Every study examined in the control theory literature emphasizes **Feedback Control** as the central mechanism enabling goal-directed behavior [Powers, 1973, 1978, **?**, Frank, 2018]. This universality across theoretical frameworks—from classical control theory to modern cybernetics—suggests that feedback represents not merely a useful modeling tool but a fundamental organizational principle underlying all goal-seeking systems.

The mathematical elegance of this framework lies in its extraordinary parsimony. A goal-directed system requires only the ability to sense its current state, compare this state to a reference condition, and generate corrective actions proportional to the detected discrepancy. This simple computational architecture can generate remarkably sophisticated behaviors while remaining mathematically tractable and empirically verifiable.

3.5.2 Hierarchical Control Architecture and Emergent Complexity

The literature reveals systematic patterns in how minimal control systems can be organized into hierarchical architectures that generate increasing behavioral complexity without requiring additional fundamental

Table 3: Minimal Agency Features in Control Theory and Cybernetics Literature

Essential Feature	Operational Definition	Functional Role	Mathematical Representation
Feedback Control	Continuous monitoring of system output relative to reference state	Enables system response to disturbances and deviations	Closed-loop transfer function
Reference Signal	Target state or desired system output	Defines goal toward which system behavior is directed	Setpoint value $r(t)$
Error Correction	Mechanism for reducing discrepancy between actual and desired states	Drives goal-seeking behavior through discrepancy reduction	Error signal $e(t) = r(t) - y(t)$
System Stability	Maintenance of bounded responses to bounded inputs	Ensures reliable goal-achievement despite perturbations	Lyapunov stability criteria

agency features [Powers, 1973, **?**, Pezzulo and Cisek, 2016]. These hierarchical organizations represent a profound insight into how sophisticated goal-directed behavior can emerge from the coordination of simple control loops.

At the lowest level, individual control systems maintain basic reference states through direct sensorimotor coupling. These systems exhibit **Goal-Directedness** in its most elementary form: the maintenance of specific physical or physiological states despite environmental perturbations. The mathematical description requires only first-order differential equations relating error signals to corrective outputs.

Higher-level control systems set the reference values for lower-level controllers, creating nested hierarchies where complex behavioral sequences emerge from the coordination of simpler goal-seeking components. This hierarchical organization enables systems to pursue abstract goals (maintaining behavioral patterns, achieving complex outcomes) while retaining the mathematical simplicity of individual control loops.

The literature demonstrates that **Memory** in control systems serves a very specific functional role, differing fundamentally from memory concepts in cognitive or computational domains [Powers, 1978, Mulder et al., 2018]. Control system memory typically consists of internal state variables that accumulate error signals over time (integral control) or predict future system behavior based on current trajectories (derivative control). This memory serves the control function directly rather than storing arbitrary information about past experiences.

3.5.3 Cybernetic Compression of Complex Dynamics

The cybernetics literature provides particularly compelling insights into how control-theoretic frameworks serve as compression strategies for modeling complex adaptive systems [**?**Mulder et al., 2018]. Cybernetic approaches treat goal-directedness as an emergent property of system organization rather than as a fundamental feature requiring explanation in terms of internal representations or planning mechanisms.

The power of cybernetic compression lies in its ability to capture essential organizational features of complex systems without requiring detailed modeling of underlying mechanisms. A biological organism maintaining homeostatic regulation can be modeled using identical mathematical frameworks as an artificial thermostat, despite vast differences in their underlying physical implementations. This abstraction enables powerful generalizations about goal-directed behavior across biological and artificial systems.

Cybernetic frameworks reveal that **System Stability** emerges as both a necessary feature for reliable goal-achievement and a sufficient condition for many forms of purposive behavior [**?**Frank, 2018]. Systems that can maintain stable reference states despite environmental perturbations exhibit functional goal-directedness regardless of their internal complexity or the sophistication of their component mechanisms.

3.5.4 Dynamical Systems Perspectives on Agency

The integration of dynamical systems theory with control-theoretic approaches provides additional insights into the mathematical foundations of minimal agency [Ijspeert et al., 2013, Carver and Scheier, 2002]. Dynamical systems frameworks treat goal-directed behavior as movement through state space toward attractor states that represent system goals.

Table 4 summarizes the key differences between traditional control theory and dynamical systems approaches to modeling goal-directed behavior.

Table 4: Control Theory vs. Dynamical Systems Approaches to Goal-Directed Behavior

Modeling Aspect	**Control Theory Approach**	**Dynamical Systems Approach**
Goal Representation	Explicit reference signals and setpoints	Attractor states in state space
Error Correction	Feedback loops with proportional-integral-derivative control	Gradient descent on potential landscapes
System Behavior	Trajectory tracking and disturbance rejection	Movement through state space toward attractors
Stability Analysis	Transfer function poles and Lyapunov methods	Basin of attraction and structural stability
Adaptation Mechanisms	Parameter adjustment and gain scheduling	Landscape modification and attractor dynamics

The dynamical systems perspective reveals that goal-directed behavior can emerge from the intrinsic dynamics of the system itself rather than requiring explicit control mechanisms. Attractor dynamics provide a mathematical framework for understanding how systems can exhibit purposive behavior through their natural tendency to evolve toward stable states, even in the absence of explicit feedback controllers.

This perspective suggests that **Goal-Directedness** might be understood as an emergent property of certain classes of dynamical systems rather than as a feature requiring special explanation. Systems with appropriate attractor landscapes naturally evolve toward goal states without requiring explicit representation of desired outcomes or sophisticated planning mechanisms.

3.5.5 Comparative Analysis: Sufficiency vs. Necessity

The control theory literature provides systematic evidence for when minimal agency features prove sufficient versus when additional capabilities become necessary for effective goal-directed behavior [**?**Pezzulo and Cisek, 2016, Mulder et al., 2018]. This comparative analysis offers crucial insights into the boundary conditions of minimal agency approaches.

For systems operating in stable environments with well-defined goals, basic feedback control proves both necessary and sufficient for effective goal achievement. Simple thermostatic control, biological homeostasis, and basic robotic navigation tasks can be accomplished using minimal control architectures without requiring sophisticated cognitive capabilities.

However, the literature identifies several contexts where additional features become necessary. **Learning and Adaptation** mechanisms are required when system parameters change over time or when environmental conditions vary beyond the range anticipated by fixed control parameters [Mulder et al., 2018, Ijspeert et al., 2013]. These adaptive mechanisms typically involve parameter adjustment algorithms that modify control system behavior based on performance feedback.

Time-varying environments create particular challenges for minimal control systems, as fixed reference signals and controller parameters may become inappropriate as conditions change. The literature suggests that hierarchical control architectures provide one solution to this challenge, with higher-level controllers adjusting the reference signals and parameters of lower-level systems based on longer-term performance criteria.

3.5.6 Boundaries and Limitations of Minimal Agency

The control theory literature provides unusually explicit discussion of the limitations and boundary conditions of minimal agency approaches [Mulder et al., 2018, **?**, Pezzulo and Cisek, 2016]. These limitations offer important insights into when additional agency features become necessary for adequate system performance.

Classical control theory approaches assume time-invariant system parameters and stationary environmental conditions. When these assumptions are violated, minimal control systems may fail to achieve their goals or may exhibit unstable behavior. The literature documents systematic failures of minimal approaches in environments requiring adaptation, learning, or coordination with other agents.

The transition from minimal agency to more sophisticated cognitive capabilities appears to occur when systems must operate in environments that cannot be adequately characterized by fixed reference signals and static control parameters. Complex goal hierarchies, temporal coordination of multiple objectives, and adaptation to novel environmental conditions seem to require additional agency features beyond basic feedback control.

Interestingly, the literature suggests that many apparent limitations of minimal control systems can be addressed through architectural modifications rather than fundamental changes to the underlying control principles. Hierarchical organization, adaptive parameter adjustment, and environmental structure can extend the range of problems addressable by control-theoretic approaches without requiring fundamentally different agency features.

3.5.7 Integration with Broader Agency Frameworks

The control theory perspective provides a crucial foundation for understanding agency across other domains examined in this paper. The mathematical precision and empirical validation available in control theory offers a solid foundation for extending agency concepts to more complex cognitive and social systems.

The fundamental insights from control theory—that goal-directed behavior emerges from feedback loops, error correction, and stability mechanisms—appear to be preserved even in sophisticated cognitive architectures. Higher-level cognitive functions can be understood as elaborate hierarchical organizations of basic control principles rather than as qualitatively different forms of agency.

This suggests that the minimal agency framework developed in control theory might serve as a foundational level for understanding more complex forms of agency in biological, cognitive, and social systems. The control-theoretic perspective provides both mathematical tools and conceptual frameworks that can be extended to domains where direct measurement and intervention are more challenging.

The systematic exploration of minimal agency in control theory thus offers not merely a specialized perspective on goal-directed behavior in engineered systems, but a foundational understanding of the mathematical principles underlying purposive action in any domain where such behavior emerges.

3.6 Cognitive Science (General)

A Cognitive Scientist might propose: "We aim to predict and explain the underlying mechanisms of human and animal cognition and behavior across a wide range of tasks. Our dominant modeling stance is to view the mind as an **Information Processing** system, often implemented through diverse computational architectures. Key features considered necessary for comprehensive models of human cognition include various **Memory Systems** (working, long-term, episodic), mechanisms for **Mental Representation** and often **Symbolic Processing**, alongside learning and reasoning capabilities [**?**Sowa, 2011]. **Bounded Rationality** and **Cognitive Constraints** (e.g., on attention, processing speed) are fundamental and necessary assumptions to explain actual performance [Cooper and Peebles, 2015]. For social behavior and communication, modeling with **Theory of Mind** is generally considered necessary. Different cognitive architectures (symbolic, connectionist, hybrid) offer competing compressions for how these features are realized and interact, and a key challenge is to bridge neural, cognitive, and behavioral levels of explanation [Doumas and Hummel, 2012, **?**]."

The cognitive science literature represents a fascinating convergence of computational metaphors, empirical observation, and theoretical synthesis. What emerges from systematic reviews of the field is a discipline wrestling with one of the most profound questions in science: how can we model the mechanisms that give rise to intelligent behavior while remaining faithful to both computational realities and empirical constraints?

3.6.1 The Information Processing Paradigm and Architectural Choices

Reviews across cognitive science consistently reveal information processing as the dominant theoretical metaphor, with researchers treating the mind as a computational system that manipulates symbolic representations [?Doumas and Hummel, 2012, Sowa, 2011]. However, this broad consensus masks important architectural debates about how such processing should be implemented.

Table 5 summarizes the major architectural approaches identified across theoretical reviews, their characteristic features, and empirical support patterns.

Table 5: Cognitive Architecture Types in Information Processing Models

Architecture Type	**Core Processing Assumptions**	**Representational Format**	**Empirical Support**
Symbolic	Rule-based manipulation of discrete symbols	Explicit symbolic structures, logical propositions	Strong for reasoning tasks, language processing
Connectionist	Parallel distributed processing across neural-like networks	Distributed activation patterns	Strong for pattern recognition, learning
Hybrid	Integration of symbolic and connectionist components	Mixed symbolic-distributed representations	Increasing preference in recent reviews
Emergent	Self-organizing systems with minimal initial structure	Dynamic, context-dependent representations	Growing support for developmental phenomena

The literature reveals a striking pattern: while symbolic approaches dominate unified theories of cognition, no review exclusively advocates purely connectionist architectures. Instead, six out of ten major reviews emphasize the necessity of hybrid models that integrate symbolic manipulation with distributed processing capabilities [Kriegeskorte and Douglas, 2018, Sowa, 2011, Doumas and Hummel, 2012].

This convergence toward hybrid architectures reflects deeper insights about the computational requirements of different cognitive tasks. **Symbolic Processing** proves necessary for tasks requiring compositional structure, logical reasoning, and explicit rule manipulation, while distributed processing becomes essential for pattern recognition, associative memory, and adaptive learning. The field has increasingly recognized that comprehensive cognitive models require both processing modes rather than treating them as competing alternatives.

3.6.2 Memory Systems as Computational Architectures

Perhaps no aspect of cognitive science demonstrates the necessity of specific agency features more clearly than the treatment of memory systems. Reviews consistently identify multiple memory systems - working memory, long-term memory, and episodic memory - as fundamental architectural components rather than optional enhancements [?Cooper and Peebles, 2015, Sowa, 2011].

The computational necessity of these memory systems emerges from the functional roles they serve in information processing:

Working Memory functions as a limited-capacity workspace for active manipulation of information, enabling complex reasoning that exceeds the scope of immediate sensory input. The literature consistently treats working memory limitations as necessary constraints for explaining human cognitive performance rather than as implementation details [Cooper and Peebles, 2015].

Long-term Memory provides the knowledge base that enables recognition, categorization, and inference processes. Reviews demonstrate that cognitive models without substantial long-term memory systems fail to capture the breadth and flexibility of human cognitive performance across domains.

Episodic Memory enables context-dependent recall and temporal reasoning about specific events. The cognitive science literature treats episodic memory as necessary for explaining how humans navigate temporally extended situations and learn from specific experiences rather than just general patterns.

The functional organization of these memory systems reveals important insights about **Computational Boundedness** in cognitive modeling. Rather than treating memory limitations as constraints to overcome, cognitive science has increasingly recognized them as design features that enable efficient information processing under realistic resource constraints.

3.6.3 Bounded Rationality as Foundational Principle

The cognitive science literature demonstrates a remarkable convergence around bounded rationality as a fundamental modeling assumption rather than a limitation to explain away [Cooper and Peebles, 2015, Kriegeskorte and Douglas, 2018]. This represents a sophisticated understanding that cognitive constraints serve computational functions rather than merely limiting performance.

Table 6 summarizes the major cognitive constraints identified as necessary features across different cognitive architectures.

Table 6: Cognitive Constraints and Their Functional Roles in Information Processing

Constraint Type	**Functional Role**	**Computational Benefit**	**Modeling Necessity**
Attention limitations	Focus computational resources on relevant information	Prevents information overload, enables selective processing	Necessary for realistic models
Processing speed limits	Constrains real-time decision making	Forces efficient algorithms, enables time-bounded decisions	Necessary for temporal behavior
Working memory capacity	Limits simultaneous information manipulation	Prevents combinatorial explosion, forces hierarchical organization	Necessary for explaining cognitive performance
Retrieval limitations	Constrains access to long-term memory	Enables context-dependent recall, prevents interference	Necessary for memory modeling

The literature reveals that these constraints are not arbitrary limitations but rather design principles that enable efficient computation under realistic resource limitations. Cognitive models that ignore these constraints systematically fail to capture human performance patterns, while models that incorporate them naturally generate human-like behavior across diverse tasks.

3.6.4 Theory of Mind and Social Cognitive Architecture

The treatment of **Theory of Mind** in cognitive science literature reveals a sophisticated understanding of when mental state modeling becomes necessary versus sufficient for explaining social behavior. Unlike some domains where theory of mind is treated as an optional enhancement, cognitive science consistently identifies it as necessary for modeling human social cognition and communication [**?**].

However, the literature demonstrates important nuances in how theory of mind is implemented across different cognitive tasks:

Basic Social Interaction may be supported by simpler mechanisms such as emotion recognition, imitation, and social learning without requiring explicit mental state attribution. The cognitive science literature suggests these mechanisms prove sufficient for many forms of social coordination.

Complex Communication requires explicit modeling of others' knowledge states, beliefs, and intentions. The literature consistently demonstrates that sophisticated language use and collaborative problem-solving require theory of mind capabilities as necessary features.

Strategic Social Behavior involving deception, cooperation in competitive contexts, and coalition formation requires recursive mental state modeling where agents must reason about others' reasoning about their own mental states.

The developmental literature within cognitive science provides particularly compelling evidence for the necessity of theory of mind. Studies consistently demonstrate that children's acquisition of theory of mind capabilities coincides with qualitative improvements in communication, social learning, and collaborative behavior, suggesting these capabilities are necessary rather than merely beneficial for sophisticated social cognition.

3.6.5 Levels of Explanation and Integration Challenges

One of the most intellectually challenging aspects of cognitive science concerns the systematic integration of explanations across neural, cognitive, and behavioral levels. Reviews consistently identify this as a central theoretical challenge rather than a mere methodological preference [Kriegeskorte and Douglas, 2018, Cooper and Peebles, 2015, **?**].

The literature reveals systematic patterns in how different levels of explanation relate to agency modeling:

Neural Level explanations focus on the implementation mechanisms that realize cognitive processes. These explanations become necessary when cognitive models must account for biological constraints, developmental patterns, or neurological disorders.

Cognitive Level explanations focus on the computational algorithms and representational structures that implement intelligent behavior. The literature consistently treats this as the primary level for understanding agency features such as memory, reasoning, and learning.

Behavioral Level explanations focus on observable patterns of performance across tasks and environments. These explanations provide the empirical constraints that cognitive models must satisfy but typically require higher-level explanations to account for underlying mechanisms.

The integration challenge emerges from the need to maintain consistency across these levels while preserving the explanatory power available at each level. Cognitive science reviews consistently argue that adequate explanations require bridging these levels rather than reducing higher levels to lower ones or treating them as independent.

3.6.6 Representation and the Symbol Grounding Problem

The cognitive science literature reveals deep engagement with questions about how internal representations relate to external reality and how symbolic structures acquire meaning through embodied interaction. This focus on representation distinguishes cognitive science from other domains that treat representational issues as implementation details.

Reviews consistently identify several key challenges for representational approaches:

Symbol Grounding concerns how abstract symbolic representations acquire meaning through connections to perceptual and motor systems. The literature demonstrates that purely symbolic approaches struggle to explain how symbols acquire semantic content, while purely distributed approaches struggle to explain compositional structure.

Representational Format questions concern whether cognitive representations are primarily symbolic, distributed, or involve hybrid formats that combine both. The literature has increasingly converged on hybrid approaches that capture benefits of both representational strategies.

Content Addressability concerns how cognitive systems retrieve relevant information from memory based on content rather than storage location. Reviews consistently identify this as a necessary feature for flexible cognitive behavior but acknowledge ongoing challenges in implementation.

The symbol grounding problem has driven cognitive science toward embodied and enactive approaches that emphasize the role of sensorimotor experience in shaping cognitive representations. However, the literature demonstrates that these approaches must still account for abstract reasoning and language comprehension that appear to transcend immediate sensorimotor experience.

3.6.7 Synthesis: Information Processing as Modeling Stance

The cognitive science literature reveals information processing not merely as a theoretical metaphor but as a systematic modeling stance that determines which agency features prove necessary versus sufficient for explaining intelligent behavior. This stance treats the mind as a computational system that processes symbolic information under realistic resource constraints.

The systematic patterns identified across reviews suggest several key insights about agency modeling in cognitive science:

Multi-level Integration emerges as both a theoretical goal and a methodological necessity. Adequate cognitive models must bridge implementation, algorithmic, and behavioral levels rather than reducing explanations to any single level.

Hybrid Architectures prove necessary for capturing the full range of cognitive phenomena. Pure symbolic or pure connectionist approaches consistently fail to account for the breadth of human cognitive capabilities.

Bounded Rationality functions as a design principle rather than a limitation. Cognitive constraints enable efficient computation under realistic resource limitations while generating characteristic patterns of human performance.

Memory Architecture provides the foundation for complex cognitive behavior. Multiple memory systems with different functional characteristics prove necessary for flexible, context-appropriate behavior.

These insights position cognitive science as providing both theoretical frameworks and empirical constraints for understanding agency more broadly. The field's systematic exploration of information processing under realistic constraints offers crucial insights into how sophisticated intelligent behavior can emerge from computational systems operating within bounded resources.

The cognitive science perspective thus offers not merely another domain-specific approach to agency modeling, but rather a foundational understanding of how computational approaches to intelligence must be structured to capture the sophisticated yet bounded nature of realistic cognitive systems.

4 Synthesis: Logical Structure of Agency Feature Dependencies

The cross-domain analysis presented in Table 2 reveals systematic patterns in the logical dependencies between agency features that transcend disciplinary boundaries. When properly interpreted through the lens of logical sufficiency and necessity, these patterns illuminate the fundamental computational architecture underlying intelligent behavior across scales and domains.

4.1 Logical Dependencies and Sufficient Conditions: Clarifying the Framework

Our analysis requires careful attention to the logical structure of sufficient and necessary conditions. A feature is **sufficient** for agency modeling in a domain if its presence logically entails the presence of all other features required for adequate predictive performance in that domain. A feature is **necessary** if it must be present for any adequate model of agency within that domain's predictive scope.

This logical interpretation reveals that our cross-domain patterns reflect underlying dependency structures between computational capabilities rather than mere empirical associations.

4.1.1 Theory of Mind as Potential Sufficient Condition

The pattern observed in cognitive science, where **Theory of Mind** appears necessary for social cognition, suggests investigating whether Theory of Mind might function as a sufficient condition for comprehensive agency modeling within this domain. If Theory of Mind logically entails other required capabilities, then its presence would guarantee the presence of all necessary agency features.

From a computational perspective, sophisticated Theory of Mind requires maintaining recursive models of others' beliefs, desires, and reasoning processes. This capability logically presupposes several foundational capacities:

- **Strategic Reasoning**: To model others as reasoning agents, a system must itself possess strategic reasoning capabilities - **Memory**: Recursive belief modeling requires maintaining historical information

about others' past actions and inferred mental states - **Goal-Directedness**: Modeling others' intentions requires understanding goal-directed behavior, which presupposes possessing goal-directedness

However, the universal necessity of **Computational Boundedness** across all domains suggests that even sophisticated Theory of Mind operates under resource constraints. This raises a crucial question: does Theory of Mind logically entail computational boundedness, or do these represent independent logical requirements?

The answer appears domain-dependent. In cognitive science, Theory of Mind as observed in natural systems inherently operates under biological computational constraints, making boundedness an implicit component of the capability. However, in artificial systems, one could theoretically implement unbounded recursive reasoning, suggesting that computational boundedness represents an additional, independent constraint.

4.2 Hierarchical Logical Structure: Dependencies Across Capability Levels

Analysis of the necessity/sufficiency patterns reveals a hierarchical logical structure where higher-level capabilities logically depend on lower-level foundations.

4.2.1 Foundational Capabilities as Universal Necessities

The universal necessity of **Goal-Directedness** and **Computational Boundedness** across all domains suggests these represent foundational logical requirements for any coherent notion of agency. These capabilities appear to function as logical primitives that cannot be reduced to combinations of other agency features.

Goal-Directedness represents the fundamental requirement that intelligent systems exhibit systematic bias toward specific outcomes rather than random behavior. Without goal-directedness, no coherent notion of agency is possible, as systems would lack the orientational structure that defines intelligent behavior.

Computational Boundedness represents the constraint that all physical implementations of intelligence operate under finite resource limitations. This appears as a necessary condition because unbounded computational capabilities would eliminate the optimization pressures that generate the architectural tradeoffs observed across domains.

4.2.2 Memory as Conditional Necessity

The pattern for **Memory** reveals conditional logical structure: memory becomes necessary when environmental dynamics exceed immediate observational capacity, but remains merely sufficient when environmental structure enables adequate performance through reactive mechanisms.

This conditional necessity reflects information-theoretic constraints. In fully observable Markov environments, optimal policies can be expressed as direct mappings from current observations to actions, making memory logically unnecessary. In partially observable environments, optimal behavior requires maintaining belief states over hidden variables, making memory logically necessary for optimal performance.

The logical structure here follows from the mathematical requirements of optimal decision-making under different observational constraints rather than from empirical associations between domain characteristics and modeling choices.

4.2.3 Strategic Reasoning and Social Cognitive Capabilities

Strategic Reasoning exhibits a similar conditional logical structure, becoming necessary when environmental dynamics depend on other optimizing agents but remaining sufficient when simpler reactive mechanisms achieve adequate coordination.

The logical dependency becomes clear through game-theoretic analysis: in single-agent environments, optimal behavior requires only optimization over environmental responses. In multi-agent environments with strategic interdependence, optimal behavior requires reasoning about others' reasoning processes, making strategic reasoning logically necessary rather than merely empirically useful.

Theory of Mind represents a higher-order capability that logically presupposes strategic reasoning but adds recursive depth. Basic strategic reasoning involves predicting others' actions based on observable

patterns. Theory of Mind involves explicitly modeling others' beliefs and reasoning processes, enabling prediction in novel situations where behavioral patterns provide insufficient information.

4.3 Domain-Specific Logical Architectures

Different domains exhibit distinct logical architectures that reflect their specific predictive requirements and computational constraints.

4.3.1 Minimal Sufficiency in Control Systems

Control systems demonstrate perhaps the most minimal sufficient condition set, where **Feedback Control**, **Goal-Directedness**, and **Computational Boundedness** appear jointly sufficient for the domain's predictive requirements. This reflects the mathematical structure of control theory, where goal-directed behavior emerges from the interaction between reference signals, error detection, and corrective actions.

The logical sufficiency here stems from the domain's focus on maintaining specific state variables despite perturbations. More complex capabilities like strategic reasoning or theory of mind become logically unnecessary because the environment can be adequately modeled as a system of differential equations without recursive social dynamics.

4.3.2 Cognitive Architecture Requirements

Cognitive science exhibits the most complex logical architecture, where **Memory Systems**, **Theory of Mind**, **Strategic Reasoning**, and **Bounded Rationality** all appear necessary for comprehensive modeling of human cognition. This complexity reflects the domain's requirement to explain behavior across diverse social and non-social contexts.

The logical necessity of multiple capabilities stems from the empirical observation that human cognition exhibits sophisticated temporal integration (requiring memory), complex social reasoning (requiring theory of mind), and systematic biases reflecting computational limitations (requiring bounded rationality). No single capability appears sufficient because human behavior exhibits irreducible complexity across multiple dimensions.

4.3.3 Evolutionary Minimal Agency

Evolutionary systems reveal an interesting logical structure where **Minimal Agency** appears sufficient for explaining population-level dynamics, with more sophisticated capabilities becoming necessary only for specific phenomena like cultural transmission.

This reflects the mathematical structure of replicator dynamics, where complex population behaviors emerge from simple individual-level rules. Strategic reasoning and theory of mind become logically unnecessary because selection pressures operate on behavioral outcomes rather than cognitive processes, enabling sophisticated collective dynamics without sophisticated individual cognition.

4.4 Formal Logical Framework for Agency Dependencies

The patterns suggest the possibility of developing a formal logical framework for characterizing agency feature dependencies across domains.

4.4.1 Logical Entailment Structure

Let us define agency features as logical predicates and explore their entailment relationships:

- **Goal-Directedness** (G): System exhibits systematic bias toward specific outcomes
- **Computational Boundedness** (B): System operates under finite resource constraints
- **Memory** (M): System maintains internal state across temporal periods
- **Strategic Reasoning** (S): System optimizes over others' decision processes

- **Theory of Mind** (T): System explicitly models others' mental states

The logical dependencies can be formalized as:

$$T \to S \quad \text{(Theory of Mind entails Strategic Reasoning)} \tag{5}$$
$$T \to M \quad \text{(Theory of Mind entails Memory for belief tracking)} \tag{6}$$
$$S \to G \quad \text{(Strategic Reasoning entails Goal-Directedness)} \tag{7}$$
$$M \to G \quad \text{(Memory entails Goal-Directedness for temporal optimization)} \tag{8}$$

4.4.2 Domain-Specific Sufficiency Conditions

Different domains can be characterized by their specific sufficiency conditions:

$$\text{Control Systems:} \quad G \land B \to \text{Adequate Agency Modeling} \tag{9}$$
$$\text{Evolutionary Systems:} \quad G \land B \land \text{Learning Rules} \to \text{Adequate Population Dynamics} \tag{10}$$
$$\text{Cognitive Science:} \quad G \land B \land M \land S \land T \to \text{Comprehensive Cognitive Modeling} \tag{11}$$

4.5 Implications for Unified Agency Theory

The logical analysis reveals that agency modeling across domains reflects systematic computational principles rather than arbitrary disciplinary choices. The hierarchical dependency structure suggests that comprehensive agency emerges through the compositional combination of foundational capabilities rather than through qualitatively distinct mechanisms.

This insight points toward the possibility of developing a compositional theory of agency where complex capabilities emerge through logical combination of simpler primitives. Such a theory could provide principled foundations for designing artificial agents and understanding natural intelligence across scales from cellular networks to social institutions.

The logical framework also suggests that apparent disagreements between domains about agency requirements often reflect different computational environments rather than fundamental theoretical differences. By making explicit the logical dependencies between capabilities and environmental requirements, we can develop more precise theories that predict optimal agency architectures for specific computational contexts.

5 Conclusion: Mapping the Landscape of Agency Modeling

This paper reframed agency as a modeling stance—a cognitive compression strategy. We detailed the game-theoretic stance (Section 3) with its core model features (Figure **??**, Table **??**). We then broadened this view by situating various scientific domains within a landscape of agency modeling stances (Section 4, Table **??**).

The "necessity" or "sufficiency" of features is relative to predictive goals. Our exploration is an initial step towards a "mathematical skeleton" of agency. True richness will come from understanding multiple modeling stances and incorporating scalar dimensions (Section 5.2). This endeavor is crucial for AI, institutional design, and understanding life.

References

Christoph Adami, Jory Schossau, and Arend Hintze. Evolutionary game theory using agent-based methods. *Physics of Life Reviews*, 19:1–26, 12 2016. ISSN 1571-0645. doi: 10.1016/j.plrev.2016.08.015. URL `http://dx.doi.org/10.1016/j.plrev.2016.08.015`.

Larbi Alaoui and Antonio Penta. Endogenous Depth of Reasoning. *The Review of Economic Studies*, 83 (4):1297–1333, oct 29 2015. ISSN 0034-6527. doi: 10.1093/restud/rdv052. URL `http://dx.doi.org/10.1093/RESTUD/RDV052`.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Man'e. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

John R Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, 2007.

Kenichi Aoki and Marcus W. Feldman. Evolution of learning strategies in temporally and spatially variable environments: A review of theory. *Theoretical Population Biology*, 91:3–19, 2 2014. ISSN 0040-5809. doi: 10.1016/j.tpb.2013.10.004. URL http://dx.doi.org/10.1016/j.tpb.2013.10.004.

Mark Broom and Chris Cannings. Evolutionary Game Theory. *Encyclopedia of Life Sciences*, nov 15 2010. doi: 10.1002/9780470015902.a0005457.pub2. URL http://dx.doi.org/10.1002/9780470015902.A0005457.PUB2.

Joanna Bryson. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(2):165–189, 4 2000. ISSN 0952-813X. doi: 10.1080/095281300409829. URL http://dx.doi.org/10.1080/095281300409829.

C. F. Camerer, T.-H. Ho, and J.-K. Chong. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, 119(3):861–898, aug 1 2004. ISSN 0033-5533. doi: 10.1162/0033553041502225. URL http://dx.doi.org/10.1162/0033553041502225.

Colin Camerer, Teck Ho, and Kuan Chong. Models of Thinking, Learning, and Teaching in Games. *American Economic Review*, 93(2):192–195, apr 1 2003. ISSN 0002-8282. doi: 10.1257/000282803321947038. URL http://dx.doi.org/10.1257/000282803321947038.

Colin F Camerer. Progress in Behavioral Game Theory. *Journal of Economic Perspectives*, 11(4):167–188, nov 1 1997. ISSN 0895-3309. doi: 10.1257/jep.11.4.167. URL http://dx.doi.org/10.1257/JEP.11.4.167.

Charles S. Carver and Michael F. Scheier. Control Processes and Self-Organization as Complementary Principles Underlying Behavior. *Personality and Social Psychology Review*, 6(4):304–315, 11 2002. ISSN 1088-8683. doi: 10.1207/s15327957pspr0604_05. URL http://dx.doi.org/10.1207/S15327957PSPR0604_05.

Claudio Collinet and Thomas Lecuit. Programmed and self-organized flow of information during morphogenesis. *Nature Reviews Molecular Cell Biology*, 22(4):245–265, jan 22 2021. ISSN 1471-0072. doi: 10.1038/s41580-020-00318-6. URL http://dx.doi.org/10.1038/s41580-020-00318-6.

Richard P. Cooper and David Peebles. Beyond SingleLevel Accounts: The Role of Cognitive Architectures in Cognitive Scientific Explanation. *Topics in Cognitive Science*, 7(2):243–258, feb 28 2015. ISSN 1756-8757. doi: 10.1111/tops.12132. URL http://dx.doi.org/10.1111/tops.12132.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2nd edition, 2006.

Vincent P Crawford, Miguel A Costa-Gomes, and Nagore Iriberri. Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications. *Journal of Economic Literature*, 51(1):5–62, mar 1 2013. ISSN 0022-0515. doi: 10.1257/jel.51.1.5. URL http://dx.doi.org/10.1257/JEL.51.1.5.

Abram Demski. Factored sets: A framework for modeling agent boundaries. *AI Alignment Forum*, 2018.

Daniel C Dennett. *The intentional stance*. MIT press, 1987.

Leonidas A. A. Doumas and John E. Hummel. *Computational Models of Higher Cognition*, pages 52–66. Oxford University Press, nov 21 2012. ISBN 0199734682. doi: 10.1093/oxfordhb/9780199734689.013.0005. URL http://dx.doi.org/10.1093/OXFORDHB/9780199734689.013.0005.

W. Tecumseh Fitch. Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, 11(3):329–364, 9 2014. ISSN 1571-0645. doi: 10.1016/j.plrev.2014.04.005. URL http://dx.doi.org/10.1016/j.plrev.2014.04.005.

Evan Flint. Agent boundaries and the problem of the boundary. *AI Alignment Forum*, 2020.

Steven A. Frank. Evolutionary design of regulatory control. I. A robust control theory analysis of tradeoffs. *bioRxiv*, may 30 2018. doi: 10.1101/332999. URL http://dx.doi.org/10.1101/332999.

Karl Friston, Michael Levin, Biswa Sengupta, and Giovanni Pezzulo. Knowing one's place: a free-energy approach to pattern regulation. *Journal of The Royal Society Interface*, 12(105):20141383, 4 2015. ISSN 1742-5689. doi: 10.1098/rsif.2014.1383. URL http://dx.doi.org/10.1098/rsif.2014.1383.

Thibaud Griessinger, Giorgio Coricelli, and Mehdi Khamassi. The computation of strategic learning in repeated social competitive interactions: Learning sophistication, reward attractor points and strategic asymmetry. *bioRxiv*, jun 13 2018. doi: 10.1101/346155. URL http://dx.doi.org/10.1101/346155.

Joseph Henrich and Robert Boyd. On Modeling Cognition and Culture: Why cultural evolution does not require replication of representations. *Journal of Cognition and Culture*, 2(2):87–112, 2002. ISSN 1567-7095. doi: 10.1163/156853702320281836. URL http://dx.doi.org/10.1163/156853702320281836.

Francis Heylighen. The meaning and origin of goal-directedness: a dynamical systems perspective. *Biological Journal of the Linnean Society*, 139(4):370–387, jun 14 2022. ISSN 0024-4066. doi: 10.1093/biolinnean/blac060. URL http://dx.doi.org/10.1093/biolinnean/blac060.

Josef Hofbauer and Karl Sigmund. Evolutionary Games and Population Dynamics. may 28 1998. doi: 10.1017/cbo9781139173179. URL http://dx.doi.org/10.1017/CBO9781139173179.

Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. *Neural Computation*, 25(2):328–373, 2 2013. ISSN 0899-7667. doi: 10.1162/neco_a_00393. URL http://dx.doi.org/10.1162/NECO_a_00393.

Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, 2011.

Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160, aug 20 2018. ISSN 1097-6256. doi: 10.1038/s41593-018-0210-5. URL http://dx.doi.org/10.1038/s41593-018-0210-5.

Marija Kuzmanovic. Behavioral influences on strategic interactions outcomes in game theory models. *Yugoslav Journal of Operations Research*, 31(1):3–22, 2021. ISSN 0354-0243. doi: 10.2298/yjor191115023k. URL http://dx.doi.org/10.2298/yjor191115023k.

Arthur D. Lander. Pattern, Growth, and Control. *Cell*, 144(6):955–969, 3 2011. ISSN 0092-8674. doi: 10.1016/j.cell.2011.03.009. URL http://dx.doi.org/10.1016/j.cell.2011.03.009.

Michael Levin. Morphogenetic fields in embryogenesis, regeneration, and cancer: non-local control of complex patterning. *Biosystems*, 109(3):243–261, 2012.

Michael Levin. Darwin's agential materials: evolutionary implications of multiscale competency in developmental biology. *Cellular and Molecular Life Sciences*, 80(6), may 8 2023. ISSN 1420-682X. doi: 10.1007/s00018-023-04790-z. URL http://dx.doi.org/10.1007/s00018-023-04790-z.

Tao Li, Yuhan Zhao, and Quanyan Zhu. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control*, 53:296–314, 2022. ISSN 1367-5788. doi: 10.1016/j.arcontrol.2022.03.003.

A.D. Mali. On the behavior-based architectures of autonomous agency. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 32(3):231–242, 8 2002. ISSN 1094-6977. doi: 10.1109/tsmcc.2002.804445. URL http://dx.doi.org/10.1109/TSMCC.2002.804445.

Santosh Manicka and Michael Levin. Minimal Developmental Computation: A Causal Network Approach to Understand Morphogenetic Pattern Formation. *Entropy*, 24(1):107, jan 10 2022. ISSN 1099-4300. doi: 10.3390/e24010107. URL http://dx.doi.org/10.3390/e24010107.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. *arXiv.org*, 2024. doi: 10.48550/ARXIV.2404.11584. URL https://arxiv.org/abs/2404.11584.

Max Mulder, Daan M. Pool, David A. Abbink, Erwin R. Boer, Peter M. T. Zaal, Frank M. Drop, Kasper van der El, and Marinus M. van Paassen. Manual Control Cybernetics: State-of-the-Art and Current Trends. *IEEE Transactions on Human-Machine Systems*, 48(5):468–485, 10 2018. ISSN 2168-2291. doi: 10.1109/thms.2017.2761342. URL http://dx.doi.org/10.1109/THMS.2017.2761342.

Martin A. Nowak, Corina E. Tarnita, and Tibor Antal. Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):19–30, jan 12 2010. ISSN 0962-8436. doi: 10.1098/rstb.2009.0215. URL http://dx.doi.org/10.1098/rstb.2009.0215.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Giovanni Pezzulo and Paul Cisek. Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Sciences*, 20(6):414–424, 6 2016. ISSN 1364-6613. doi: 10.1016/j.tics.2016.03.013. URL http://dx.doi.org/10.1016/j.tics.2016.03.013.

Giovanni Pezzulo and Michael Levin. Top-down models in biology: explanation and control of complex living systems above the molecular level. *Journal of The Royal Society Interface*, 13(124):20160555, 11 2016. ISSN 1742-5689. doi: 10.1098/rsif.2016.0555. URL http://dx.doi.org/10.1098/rsif.2016.0555.

Léo Pio-Lopez, Johanna Bischof, Jennifer V. LaPalme, and Michael Levin. The scaling of goals from cellular to anatomical homeostasis: an evolutionary simulation, experiment and analysis. *Interface Focus*, 13(3), apr 14 2023. ISSN 2042-8901. doi: 10.1098/rsfs.2022.0072. URL http://dx.doi.org/10.1098/rsfs.2022.0072.

William T. Powers. Feedback: Beyond Behaviorism. *Science*, 179(4071):351–356, jan 26 1973. ISSN 0036-8075. doi: 10.1126/science.179.4071.351. URL http://dx.doi.org/10.1126/science.179.4071.351.

William T. Powers. Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review*, 85(5):417–435, 9 1978. ISSN 1939-1471. doi: 10.1037/0033-295x.85.5.417. URL http://dx.doi.org/10.1037/0033-295X.85.5.417.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Viking, 2019.

Herbert A Simon. *Models of bounded rationality*. MIT Press, 1982.

John F. Sowa. *Cognitive Architectures for Conceptual Structures*, pages 35–49. Springer Berlin Heidelberg, 2011. ISBN 9783642226878. doi: 10.1007/978-3-642-22688-5_3. URL http://dx.doi.org/10.1007/978-3-642-22688-5_3.

Malcolm S. Steinberg. Goal-directedness in embryonic development. *Integrative Biology: Issues, News, and Reviews*, 1(2):49–59, 1998. ISSN 1093-4391. doi: 10.1002/(sici)1520-6602(1998)1:2⟨49::aid-inbi3⟩3.0.co;2-z. URL http://dx.doi.org/10.1002/(SICI)1520-6602(1998)1:2<49::AID-INBI3>3.0.CO;2-Z.

Sonia E. Sultan, Armin P. Moczek, and Denis Walsh. Bridging the explanatory gaps: What can we learn from a biological agency perspective? *BioEssays*, 44(1), nov 7 2021. ISSN 0265-9247. doi: 10.1002/bies.202100185. URL http://dx.doi.org/10.1002/bies.202100185.

F.M. Toates and J. Archer. A comparative review of motivational systems using classical control theory. *Animal Behaviour*, 26:368–380, 5 1978. ISSN 0003-3472. doi: 10.1016/0003-3472(78)90055-6. URL http://dx.doi.org/10.1016/0003-3472(78)90055-6.

David Vernon, Giorgio Metta, and Giulio Sandini. A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. *IEEE Transactions on Evolutionary Computation*, 11(2):151–180, 4 2007. ISSN 1089-778X. doi: 10.1109/tevc.2006.890274. URL http://dx.doi.org/10.1109/TEVC.2006.890274.

John Wentworth. Natural abstractions: Key claims, theorems, and critiques. *AI Alignment Forum*, 2022.

Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 6 1995. ISSN 0269-8889. doi: 10.1017/s0269888900008122. URL http://dx.doi.org/10.1017/S0269888900008122.

James R. Wright and Kevin Leyton-Brown. Predicting human behavior in unrepeated, simultaneous-move games. *Games and Economic Behavior*, 106:16–37, 11 2017. ISSN 0899-8256. doi: 10.1016/j.geb.2017.09.009. URL http://dx.doi.org/10.1016/j.geb.2017.09.009.