A Natural History of Agency

Jonas Hallgren

October 2025

1 Introduction

2 The Evolutionary Foundations of Agency Modeling: A Multi-Scale Dynamical Systems Perspective

2.1 Axiomatic Assumptions: Evolutionary Landscapes and Free Energy Minimization

Before we can meaningfully compare agency modeling strategies across disciplines, we must state our foundational assumptions explicitly. Our framework rests on two interconnected axioms that emerge from the deepest insights of modern theoretical biology and physics of complex systems.

Axiom 1: Evolutionary Optimization. All systems we model as agents exist within evolutionary landscapes, where survival and persistence depend on effective navigation of environmental challenges over time. This is not merely true for biological organisms, but extends to economic institutions [?], cognitive representations [?], artificial systems [?], and even scientific theories themselves [?]. The agency we observe is thus always agency-under-selection, shaped by the relentless pressure to persist in competitive environments.

Axiom 2: Variational Free Energy Minimization. The second axiom follows necessarily from the first through a fundamental insight: evolutionary fitness is determined by predictive accuracy [??]. Systems that survive are those that make better predictions about their environment, enabling more effective actions and resource allocation. Since predictive accuracy can be measured as the inverse of variational free energy [?], evolutionary selection necessarily drives systems toward free energy minimization.

We therefore assume that persistent systems minimize a variational free energy functional:

$$\mathcal{F}[q,t] = \int q(\mathbf{s},t)\,[\log q(\mathbf{s},t) - \log p(\mathbf{o}(t),\mathbf{s})]\,d\mathbf{s} \tag{1}$$

where $q(\mathbf{s},t)$ represents the system's belief distribution over hidden states $\mathbf{s}$ at time t, and $p(\mathbf{o}(t),\mathbf{s})$ represents the generative model linking observations $\mathbf{o}(t)$ to hidden states. This functional captures the fundamental trade-off between maintaining coherent internal models and accurately predicting sensory evidence.

The logical chain is straightforward: evolutionary landscapes select for fitness → fitness equals predictive accuracy → predictive accuracy equals free energy minimization → therefore, persistent systems minimize free energy. Detailed proofs of this relationship can be found in **??**, but the core insight is that agency emerges as an inevitable consequence of evolutionary optimization under uncertainty.

These axioms provide the bedrock upon which all agency modeling must rest. They tell us that agency is not a mysterious emergent property, but a mathematically precise consequence of evolutionary optimization under uncertainty. The systems we study have survived precisely because they became effective at minimizing prediction error while maintaining tractable internal representations—they became, in essence, evolved compression engines.

2.2 The Historical Development of Scale-Specific Modeling Stances

The intellectual history of agency modeling can be understood as a progressive exploration of different scales within coupled hierarchical systems, each scale requiring its own compression strategy to remain tractable while preserving predictive power.

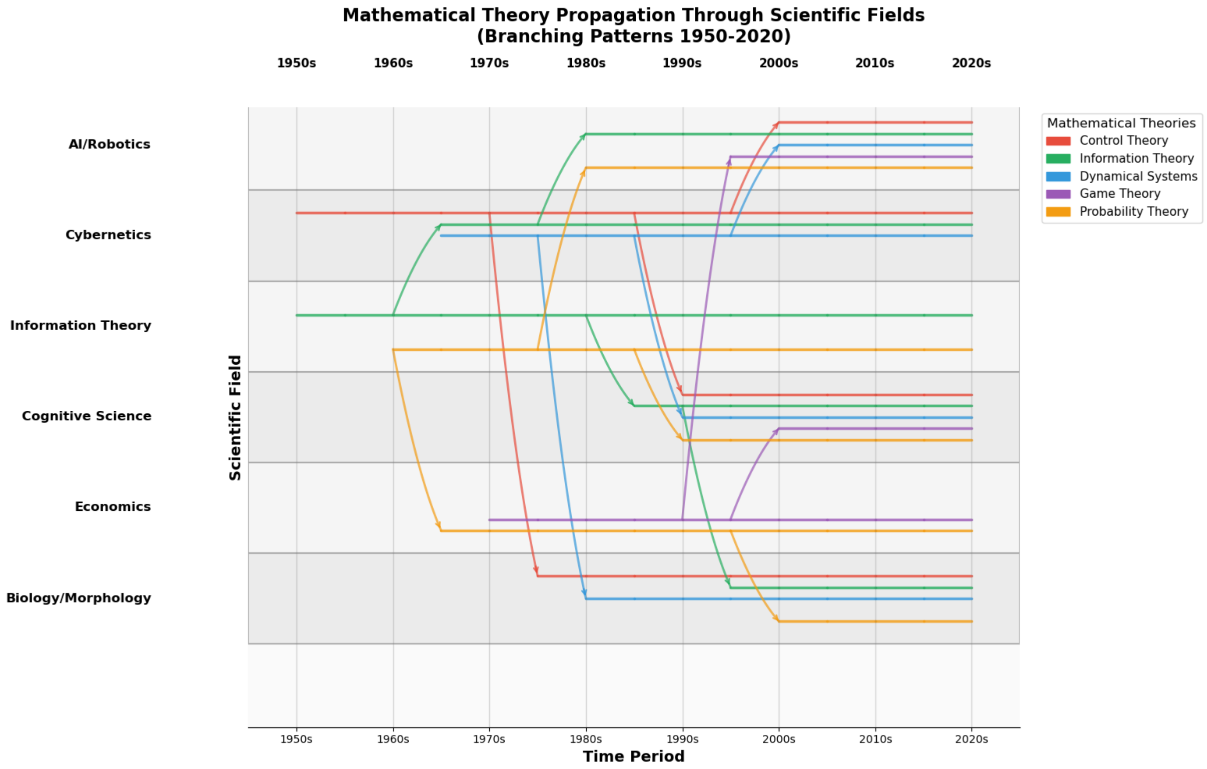

Figure 1: The evolution of the underlying scientific fields that we're analysing for perspectives of agency.

2.2.1 The Cybernetic Foundation: Control and Feedback (1940s-1960s)

The earliest systematic approach to agency modeling emerged from cybernetics [?], which discovered that goal-directed behavior could be understood through feedback control principles. This represented the first major compression strategy: treat complex systems as black boxes characterized by their input-output relationships and feedback dynamics.

The cybernetic stance focuses on minimal structural assumptions—systems as transfer functions mapping inputs to outputs—while emphasizing functional dynamics through control theory. The free energy minimization appears here in its most elementary form: maintaining system stability through error correction. For a simple feedback controller:

$$u(t) = -K[r(t) - y(t)] \tag{2}$$

where $u(t)$ is the control signal, $r(t)$ the reference signal, $y(t)$ the system output, and K the feedback gain. This implements a primitive form of free energy minimization by continuously reducing the prediction error $e(t) = r(t) - y(t)$.

This stance proved remarkably successful for engineering applications but revealed its limitations when applied to systems exhibiting learning, adaptation, or strategic interaction. The cybernetic compression sacrifices internal complexity for analytical tractability, making it ideal for understanding homeostatic regulation but inadequate for modeling systems that must model other systems.

2.2.2 The Biological Scale: Developmental and Evolutionary Dynamics (1960s-1980s)

As molecular biology revealed the breathtaking complexity of cellular behavior, researchers faced a new modeling challenge: how to understand systems that exhibit apparent goal-directedness without centralized control. Developmental biology pioneered the *teleological stance* [?], which compresses distributed molecular networks into goal-directed cellular and tissue behaviors.

This biological stance emphasizes structural organization—gene regulatory networks, morphogenetic fields, signaling pathways—while treating functional dynamics through teleological descriptions. The free energy principle manifests here through what we might call "morphogenetic free energy": cells collectively minimize the discrepancy between their current state and their target morphological configuration [?].

Consider a developing tissue where cells must coordinate to form a specific anatomical structure. The system can be modeled as minimizing:

$$\mathcal{F}_{\text{morpho}}[c,t] = \sum_i \left[\mathcal{H}[c_i(t)] + \mathcal{D}[c_i(t), \phi_{\text{target}}]\right] \tag{3}$$

where $c_i(t)$ represents the state of cell i at time t, $\mathcal{H}[c_i(t)]$ captures the entropy of cellular state uncertainty, and $\mathcal{D}[c_i(t), \phi_{\text{target}}]$ measures the divergence from the target morphological field ϕ_{target}.

The biological stance's great insight was recognizing that complex coordination could emerge from local interactions without global controllers—a compression that proved essential for understanding everything from embryogenesis to ecosystem dynamics. However, this stance struggles with systems that require explicit modeling of other agents' mental states or strategic reasoning about competitive interactions.

2.2.3 The Economic Scale: Strategic Interaction and Bounded Rationality (1970s-1990s)

The emergence of behavioral economics represented a revolutionary compression innovation: the *bounded rationality stance* [?]. Faced with systematic failures of classical rational choice theory, economists developed models that compress cognitive complexity into tractable heuristics while preserving strategic reasoning capabilities.

This economic stance balances structural assumptions (utility functions, choice sets, information structures) with functional dynamics (learning rules, adaptation mechanisms, strategic updating). The free energy principle appears here as bounded optimization: agents minimize prediction error about others' behavior while operating under cognitive resource constraints.

For a boundedly rational agent in a strategic environment:

$$\mathcal{F}_{\text{strategic}}[b,t] = \mathbb{E}_{b(\theta_{-i})}\left[\mathcal{L}[\pi_i(b), \theta_{-i}]\right] + \lambda \mathcal{C}[\pi_i] \tag{4}$$

where $b(\theta_{-i})$ represents beliefs about others' types θ_{-i}, $\mathcal{L}[\pi_i(b), \theta_{-i}]$ captures the loss from mismatched strategies, and $\mathcal{C}[\pi_i]$ penalizes cognitive complexity in strategy computation.

The economic stance's breakthrough was realizing that rationality itself must be understood as resource-bounded, leading to level-k reasoning models [?] and other tractable approximations to full strategic reasoning. This compression proved essential for understanding market dynamics and social coordination, though it often abstracts away the detailed cognitive mechanisms underlying decision-making.

2.2.4 The Cognitive Scale: Mechanistic Architecture and Social Reasoning (1980s-2000s)

Cognitive science faced a unique challenge: explaining how biological systems implement the computational functions that economics and biology simply assumed. This led to the *mechanistic stance*, which compresses psychological phenomena into explicit computational architectures with detailed processing mechanisms.

The cognitive stance emphasizes both structural organization (memory systems, attention mechanisms, representational formats) and functional dynamics (learning algorithms, reasoning processes, social cognition). The free energy principle manifests through predictive processing frameworks [?], where brains minimize prediction error through hierarchical message passing:

$$\mathcal{F}_{\text{cognitive}}[h,t] = \sum_l \mathbb{E}_{q(h^l)}\left[\epsilon^l(h^l, h^{l+1})\right] + \mathcal{D}[q(h^l) \parallel p(h^l)] \tag{5}$$

where h^l represents hidden states at hierarchical level l, ϵ^l captures prediction errors between levels, and the divergence term regularizes belief complexity.

The cognitive stance's crucial innovation was explicitly modeling Theory of Mind as recursive belief attribution [?], enabling predictions about social behavior that neither cybernetic nor biological stances could capture. However, this mechanistic detail often makes cognitive models computationally intractable for large-scale social or economic phenomena.

2.2.5 The Artificial Scale: Hybrid Architectures and Adaptive Systems (1990s-Present)

Artificial intelligence inherited the modeling challenges of all previous scales while adding a new constraint: systems must actually work in practice. This led to the *hybrid stance*, which combines reactive, deliberative, and learning components based on task requirements [?].

The AI stance treats both structure and function as design parameters to be optimized. The free energy principle appears here through reinforcement learning frameworks [?], where artificial agents minimize the divergence between expected and actual outcomes:

$$\mathcal{F}_{\text{AI}}[Q, \pi, t] = \mathbb{E}_\pi \left[\sum_{t'} \gamma^{t'-t} \delta_{t'} \right] + \alpha \mathcal{H}[\pi] \tag{6}$$

where Q represents value estimates, π represents the policy, $\delta_{t'}$ are temporal difference errors, and $\mathcal{H}[\pi]$ provides entropy regularization for exploration.

The AI stance's contribution has been demonstrating how different modeling approaches can be combined systematically, leading to architectures that exhibit cybernetic stability, biological adaptation, economic rationality, and cognitive flexibility within unified systems.

2.3 Cross-Scale Integration: The Emergence of Hierarchical Agency

What emerges from this historical analysis is a profound insight: the different modeling stances are not competing theories of agency, but complementary approaches to different scales of the same hierarchical dynamical system. Each scale exhibits agency through free energy minimization, but the relevant structures, timescales, and functional constraints differ dramatically.

Consider a human economic agent making a market decision. At the neural scale (cognitive stance), the brain implements predictive processing to minimize sensory prediction errors. At the psychological scale (still cognitive stance), working memory and reasoning systems minimize uncertainty about decision outcomes. At the behavioral scale (economic stance), the person implements bounded rational strategies that minimize regret in strategic interactions. At the social scale (multiple stances), markets and institutions evolve to minimize coordination failures.

Each scale can be understood as implementing its own version of free energy minimization:

$$\begin{aligned}
\text{Neural: } & \mathcal{F}_{\text{neural}}[q_{\text{neural}}, t] \qquad (7) \\
\text{Psychological: } & \mathcal{F}_{\text{psych}}[q_{\text{psych}}, t] \qquad (8) \\
\text{Behavioral: } & \mathcal{F}_{\text{behav}}[q_{\text{behav}}, t] \qquad (9) \\
\text{Social: } & \mathcal{F}_{\text{social}}[q_{\text{social}}, t] \qquad (10)
\end{aligned}$$

The genius of different modeling stances lies in their recognition that effective agency modeling requires compression strategies tailored to specific scales. The cybernetic stance compresses away everything except feedback dynamics. The biological stance compresses away individual cognition while preserving coordination. The economic stance compresses away neural mechanisms while preserving strategic reasoning. The cognitive stance preserves mechanistic detail while abstracting away lower-level biology. The AI stance treats all levels as design parameters to be optimized.

This multi-scale perspective reveals why attempts to reduce agency to any single level inevitably fail. Agency is not located at any particular scale—it emerges from the hierarchical organization of free energy minimization across multiple coupled scales. Different disciplines have developed their modeling stances not out of theoretical preference, but from the practical necessity of working effectively at different levels of this hierarchy.

2.4 The Mathematical Foundations of Agency Compression

The historical development of agency modeling across scales has crystallized into a remarkable collection of mathematical frameworks, each offering distinct lenses through which to view and compress the complexity of intelligent behavior. These frameworks are not mere analytical conveniences—they represent fundamental discoveries about how different aspects of agency can be mathematically characterized, predicted, and understood. What emerges from this mathematical landscape is a profound insight: the agency concepts that different fields employ are not arbitrary cultural constructs, but natural consequences of applying specific mathematical compression strategies to the challenge of understanding complex adaptive systems.

Each mathematical framework embeds particular assumptions about which aspects of agency are fundamental and which can be safely abstracted away. Like the legendary blind scholars examining different parts of an elephant, each mathematical approach illuminates certain features of agency while necessarily obscuring others. The genius lies not in finding the "correct" mathematical description, but in understanding how different mathematical frameworks enable different kinds of insights and predictions.

2.4.1 Dynamical Systems Theory: The Mathematics of Temporal Evolution

At the foundation of all agency modeling lies dynamical systems theory, which provides the mathematical language for describing how systems evolve through time. This framework treats agency as emerging from the temporal evolution of system states according to governing equations, whether deterministic or stochastic [**?**].

The dynamical systems perspective compresses agency into state spaces, flow fields, and attractors. An agent becomes a trajectory through state space, guided by vector fields that capture the underlying dynamics. Agency manifests through the system's capacity to move toward specific attractors (goals) while avoiding others (threats), mediated by the landscape of the state space itself.

Consider the elegant formulation of an adaptive agent as a dynamical system:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{p}, t) \tag{11}$$

where $\mathbf{x}$ represents the agent's internal state, $\mathbf{u}$ represents environmental inputs, $\mathbf{p}$ represents parameters that can adapt over time, and $\mathbf{f}$ captures the evolutionary dynamics. The profound insight of this framework is that agency emerges not from mysterious internal properties, but from the geometric structure of the state space and the flow patterns that govern transitions between states.

This mathematical lens has proven particularly powerful in developmental biology, where morphogenetic processes can be understood as dynamical systems evolving toward stable morphological attractors [**?**]. Cells follow dynamical trajectories through gene expression space, guided by regulatory networks that create attractor landscapes favoring specific developmental outcomes. The seeming purposefulness of embryonic development emerges naturally from the mathematical structure of these high-dimensional dynamical systems.

2.4.2 Control Theory: The Mathematics of Goal-Directedness

Control theory provides perhaps the most direct mathematical framework for capturing goal-directed behavior, treating agency as the capacity to regulate system behavior toward desired setpoints despite environmental disturbances [**??**]. This framework compresses agency into feedback loops, transfer functions, and stability criteria.

The control-theoretic perspective reveals agency as emerging from the mathematical relationship between error signals and corrective actions:

$$u(t) = -K \int_0^t e(\tau) d\tau - K_p e(t) - K_d \frac{de(t)}{dt} \tag{12}$$

where $e(t) = r(t) - y(t)$ represents the error between desired reference signal $r(t)$ and actual output $y(t)$, while K, K_p, and K_d represent integral, proportional, and derivative control gains. This deceptively simple equation captures a profound truth: goal-directedness can emerge from purely mechanistic processes that minimize discrepancies between desired and actual states.

The mathematical elegance of control theory lies in its demonstration that agency requires neither consciousness nor intentionality—only the appropriate coupling between perception, error detection, and action. From bacterial chemotaxis [?] to human motor control [?], the same mathematical principles govern how systems achieve goals through feedback regulation.

Yet control theory's compression strategy necessarily abstracts away the internal complexity of the controlled system, treating it as a transfer function relating inputs to outputs. This mathematical choice enables powerful engineering applications while potentially obscuring the rich internal dynamics that more detailed approaches might reveal.

2.4.3 Game Theory: The Mathematics of Strategic Interaction

Game theory emerges as the natural mathematical framework when agency must be understood in contexts involving multiple interacting decision-makers [?]. This framework compresses agency into strategies, payoffs, and equilibrium concepts, revealing how rational behavior emerges from the mathematical structure of strategic interdependence.

The game-theoretic compression treats agents as optimizers of expected utility functions:

$$\max_{s_i \in S_i} \mathbb{E}\left[u_i(s_i, s_{-i}) \mid \text{beliefs about } s_{-i}\right] \tag{13}$$

where agent i selects strategy s_i from strategy set S_i to maximize expected utility u_i, given beliefs about other agents' strategies s_{-i}. The mathematical beauty of this framework lies in how complex social phenomena—cooperation, competition, coordination, communication—emerge naturally from the solution concepts (Nash equilibria, evolutionary stable strategies, correlated equilibria) that characterize stable configurations of strategic interaction.

Game theory's compression strategy proves particularly powerful for economic analysis, where market behaviors can be understood as equilibrium outcomes of strategic games between rational agents [?]. The framework's abstraction away from psychological mechanisms enables tractable analysis of large-scale social systems while preserving the essential strategic logic that drives collective outcomes.

The recent development of algorithmic game theory has revealed deep connections between game-theoretic concepts and computational complexity theory [?], suggesting that strategic reasoning itself may be understood as a computational process subject to fundamental complexity constraints.

2.4.4 Information Theory: The Mathematics of Communication and Computation

Information theory provides a fundamental mathematical framework for understanding agency through the lens of information processing, compression, and transmission [?]. This perspective treats agents as information-processing systems that must efficiently encode, transmit, and decode signals about their environment and internal states.

The information-theoretic compression of agency centers on entropy, mutual information, and channel capacity:

$$H(X) = -\sum_x p(x) \log p(x) \tag{14}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{15}$$

$$C = \max_{p(x)} I(X;Y) \tag{16}$$

where $H(X)$ measures the uncertainty in random variable X, $I(X;Y)$ quantifies the mutual information between variables X and Y, and C represents the maximum information that can be reliably transmitted through a communication channel.

From this mathematical perspective, agency emerges as the capacity to reduce uncertainty about the environment through optimal information acquisition and processing. An agent's effectiveness becomes fundamentally limited by information-theoretic bounds on sensing, computation, and communication [?].

This framework has proven particularly powerful in neuroscience, where neural computation can be understood as implementing optimal information-theoretic strategies for encoding sensory information and coordinating behavioral responses [?]. The mathematical constraints of information theory explain why biological systems exhibit specific architectural features—sparse coding, hierarchical organization, predictive processing—that optimize information processing efficiency.

The emergence of machine learning has revealed deep connections between information theory and statistical learning, suggesting that agency itself may be understood as a form of statistical inference under computational constraints [?].

2.4.5 Probability Theory: The Mathematics of Uncertainty and Belief

Probability theory provides the mathematical framework for understanding agency under uncertainty, treating agents as Bayesian reasoners who update beliefs based on evidence and make decisions based on expected utilities [?]. This framework compresses agency into probability distributions, likelihood functions, and decision rules.

The probabilistic compression represents agency through Bayesian inference:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})} \tag{17}$$

$$\mathbf{a}^* = \arg\max_{\mathbf{a}} \mathbb{E}_{p(\theta|\mathbf{x})}[u(\mathbf{a},\theta)] \tag{18}$$

where $p(\theta|\mathbf{x})$ represents posterior beliefs about hidden state θ given observations $\mathbf{x}$, and $\mathbf{a}^*$ represents the optimal action that maximizes expected utility under uncertainty. Agency emerges from the capacity to maintain coherent probabilistic beliefs and make optimal decisions despite incomplete information.

This mathematical perspective has become increasingly central to cognitive science, where human reasoning can be understood as approximate Bayesian inference [?]. The mathematical constraints of probability theory explain characteristic features of human cognition—confirmation bias, anchoring effects, overconfidence—as rational adaptations to computational limitations in probabilistic reasoning.

2.4.6 Cross-Framework Integration: Hybrid Mathematical Approaches

The true power of mathematical agency modeling emerges when these different frameworks are combined to capture multiple aspects of complex systems simultaneously. Real agency phenomena rarely admit to purely game-theoretic or purely information-theoretic analysis—they require hybrid mathematical approaches that integrate insights from multiple frameworks.

Consider, for example, the mathematical description of a learning agent in a social environment:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\mathbf{u},\mathbf{p}) \quad \text{(dynamical systems)} \tag{19}$$

$$\mathbf{u}^* = \arg\max_{\mathbf{u}} \mathbb{E}[R|\mathbf{x},\mathbf{u}] \quad \text{(optimization)} \tag{20}$$

$$R = \sum_j u_j(s_i, s_j) \quad \text{(game theory)} \tag{21}$$

$$s_i = \arg\max_s \mathbb{E}_{p(\mathbf{s}_{-i}|\mathbf{o})}[u_i(s,\mathbf{s}_{-i})] \quad \text{(probability theory)} \tag{22}$$

$$I(\mathbf{x};\mathbf{o}) \geq I_{\min} \quad \text{(information theory)} \tag{23}$$

This hybrid mathematical representation captures how an agent's internal state evolves dynamically, makes optimal decisions under uncertainty, engages in strategic interactions with other agents, maintains probabilistic beliefs about others' strategies, and operates under information-processing constraints. No single mathematical framework could capture this full complexity—the power emerges from their systematic integration.

The mathematical relationships between these frameworks reveal deep structural connections: game theory emerges as a special case of optimization theory when objective functions depend on others' actions;

information theory constrains the complexity of strategies that can be effectively implemented; dynamical systems theory governs how probabilistic beliefs and strategic behaviors evolve over time; control theory emerges when optimization objectives involve maintaining desired system states.

This mathematical perspective transforms our understanding of the historical development described earlier. The cybernetic stance emphasized control-theoretic and dynamical systems approaches. The biological stance integrated dynamical systems with information theory and graph theory. The economic stance combined game theory with optimization theory and probability theory. The cognitive stance required hybrid approaches integrating all frameworks to capture the full complexity of human reasoning and social interaction.

The profound insight that emerges from this mathematical analysis is that agency is not a unified phenomenon that admits to a single mathematical description. Rather, agency is a multifaceted capacity that manifests differently depending on which mathematical lens we apply. The compression strategies that different fields have developed reflect their implicit choices about which mathematical frameworks best capture the aspects of agency most relevant to their explanatory purposes.

This mathematical foundation provides the conceptual scaffolding for understanding how different modeling stances emerge, evolve, and relate to each other across scales and domains. In the following analysis, we will examine how these mathematical frameworks have been deployed by different scientific disciplines to develop their characteristic approaches to agency modeling, and how the necessity or sufficiency of different agency features depends fundamentally on the mathematical frameworks being employed.