# Scalar Properties of Agency

Jonas Hallgren

October 2025

## 1 Introduction

## 2 Information-Theoretic Foundations of Scalar Agency Properties

**OBS:** (This is an expertimental section)

From the perspective of computational neuroscience and theoretical computer science, the categorical treatment of agency features in Table 1 represents a fundamental limitation that obscures the underlying continuous dynamics governing intelligent behavior. Drawing upon the mathematical frameworks of active inference, computational complexity theory, and embedded agency, we can reconceptualize these features as scalar fields operating over information-theoretic state spaces.

#### 2.0.1 Memory as Information Integration Capacity

Rather than treating Memory as a binary property, we can formalize it as a scalar field $\mathcal{M} : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^+$ where $\mathcal{S}$ represents the space of possible system states and $\mathcal{T}$ represents temporal horizons. This formalization captures the fundamental insight that memory represents the system's capacity to integrate information across temporal scales, measured in terms of mutual information preservation:

$$\mathcal{M}(s,t) = I(S_0; S_t | \text{observations}_{0:t}) \tag{1}$$

where $I(\cdot;\cdot|\cdot)$ denotes conditional mutual information. This scalar representation reveals critical phase transitions: systems exhibit qualitatively different behavioral capabilities as $\mathcal{M}$ crosses specific threshold values determined by the complexity of temporal dependencies in their environment.

The sufficiency threshold for memory can be characterized through the lens of predictive information theory. A system possesses "sufficient" memory for a given computational context when:

$$\mathcal{M}(s,t) \geq \max_{\pi \in \Pi} H(S_{t+1} | S_t, \pi) \tag{2}$$

where $\Pi$ represents the space of possible policies and $H(\cdot|\cdot)$ denotes conditional entropy. This threshold captures the minimum information integration capacity required to maintain predictive accuracy across the system's planning horizon.

#### 2.0.2 Strategic Reasoning as Recursive Depth Scaling

Strategic Reasoning emerges as a natural scalar property when viewed through the computational complexity lens. Rather than a binary capability, strategic reasoning can be formalized as a function $\mathcal{R} : \mathcal{A} \times \mathbb{N} \to [0,1]$ mapping agent architectures and recursion depths to reasoning capabilities:

$$\mathcal{R}(a,k) = \frac{1}{Z} \sum_{i=0}^{k} \exp(-\beta \cdot C_{\text{comp}}(i)) \cdot \mathcal{P}_{\text{correct}}(a,i) \tag{3}$$

where $C_{\text{comp}}(i)$ represents the computational cost of $i$-level recursive reasoning, $\mathcal{P}_{\text{correct}}(a,i)$ denotes the probability of correct strategic predictions at recursion depth $i$, and $Z$ serves as a normalization constant.

This formalization reveals that strategic reasoning exhibits smooth scaling properties with sharp phase transitions at critical recursion depths. The sufficiency threshold occurs when the marginal benefit of additional recursive depth falls below the computational cost:

$$\frac{\partial \mathcal{R}}{\partial k} = \frac{\partial C_{\text{comp}}}{\partial k} \tag{4}$$

#### 2.0.3 Theory of Mind as Belief State Dimensionality

Theory of Mind can be reconceptualized as a scalar field measuring the dimensionality of belief state representations that an agent can maintain about other agents. Following the active inference framework, we can formalize this as:

$$\mathcal{T}(a) = \log \dim(\mathcal{B}_a) \tag{5}$$

where $\mathcal{B}_a$ represents the space of belief states that agent $a$ can represent about other agents' mental states. This scalar representation captures the exponential scaling of representational complexity as theory of mind sophistication increases.

The critical insight from this perspective is that Theory of Mind exhibits threshold behavior: below a minimum dimensionality threshold $\mathcal{T}_{\min}$, agents cannot effectively model other minds; above this threshold, capabilities scale smoothly but with exponentially increasing computational costs.

#### 2.0.4 Goal-Directedness as Variational Free Energy Minimization

From the free energy principle perspective, Goal-Directedness emerges as a scalar property measuring the strength of an agent's tendency toward variational free energy minimization. We can formalize this as:

$$\mathcal{G}(a, t) = -\frac{d}{dt}\mathcal{F}[q(s_t), p(s_t|m)] \tag{6}$$

where $\mathcal{F}$ represents variational free energy, $q(s_t)$ denotes the agent's posterior beliefs about hidden states, and $p(s_t|m)$ represents the generative model. This formalization reveals that goal-directedness is not binary but represents the rate at which agents minimize prediction error and maintain preferred states.

### 2.1 Vector Field Dynamics of Agency Feature Interactions

The scalar formalization enables us to conceptualize agency features as vector fields operating over information-theoretic state spaces. The interaction dynamics between features can be represented as:

$$\frac{d\vec{\mathcal{A}}}{dt} = \mathcal{F}_{\text{interaction}}(\mathcal{M}, \mathcal{R}, \mathcal{T}, \mathcal{G}, \mathcal{B}) \tag{7}$$

where $\vec{\mathcal{A}} = [\mathcal{M}, \mathcal{R}, \mathcal{T}, \mathcal{G}]^T$ represents the agency feature vector and $\mathcal{B}$ represents computational boundedness constraints.

This vector field perspective reveals several critical insights:

1. Attractor Dynamics: Different combinations of agency features correspond to stable attractors in the vector field, explaining why certain architectural patterns emerge consistently across domains.

2. Phase Transitions: Sharp changes in vector field topology correspond to qualitative transitions in agent capabilities, such as the emergence of social cognition or temporal planning.

3. Constraint Manifolds: Computational boundedness creates constraint manifolds that limit possible trajectories through agency feature space.

#### 2.1.1 Differentiable Agency Architecture Optimization

The scalar field formalization enables the development of differentiable optimization frameworks for agency architectures. We can define an objective function over agency feature configurations:

$$\mathcal{L}(\vec{\mathcal{A}}) = \mathbb{E}_{\tau \sim \pi_{\vec{\mathcal{A}}}} \left[ \sum_{t=0}^{T} \gamma^t r(s_t, a_t) \right] - \lambda \cdot \mathcal{C}_{\text{comp}}(\vec{\mathcal{A}}) \tag{8}$$

where $\tau$ represents trajectories generated by policy $\pi_{\vec{\mathcal{A}}}$ parameterized by agency features $\vec{\mathcal{A}}$, $r(s_t, a_t)$ denotes reward, and $\mathcal{C}_{\text{comp}}$ represents computational costs.

This framework enables gradient-based optimization of agency architectures:

$$\vec{\mathcal{A}}_{t+1} = \vec{\mathcal{A}}_t + \alpha \nabla_{\vec{\mathcal{A}}} \mathcal{L}(\vec{\mathcal{A}}_t) \tag{9}$$

## 2.2 Empirical Characterization of Phase Transitions

The scalar field perspective suggests specific empirical approaches for characterizing phase transitions in agency capabilities:

### 2.2.1 Memory Capacity Thresholds

Experimental protocols can systematically vary memory constraints while measuring performance on temporal integration tasks. The critical memory threshold $\mathcal{M}_c$ can be identified through power-law scaling near the transition:

$$\mathcal{P}_{\text{success}}(\mathcal{M}) \propto |\mathcal{M} - \mathcal{M}_c|^{\beta} \tag{10}$$

### 2.2.2 Strategic Reasoning Depth Scaling

Computational experiments can measure the scaling of strategic reasoning capabilities with recursive depth constraints. The optimal depth $k^*$ emerges from the trade-off between accuracy and computational cost:

$$k^* = \arg\max_k \left[ \mathcal{P}_{\text{correct}}(k) - \lambda \cdot \exp(\gamma k) \right] \tag{11}$$

### 2.2.3 Theory of Mind Dimensionality Requirements

Virtual environment experiments can systematically vary the complexity of other agents' mental states while measuring prediction accuracy. The minimum sufficient Theory of Mind dimensionality can be characterized through information-theoretic analysis of belief state representations.

## 2.3 Cross-Domain Scaling Laws

The mathematical framework enables derivation of scaling laws that predict agency requirements across different computational domains. These scaling laws take the form:

$$\mathcal{A}_{\text{required}}(\mathcal{D}) = \mathcal{A}_0 \cdot \left( \frac{\mathcal{C}(\mathcal{D})}{\mathcal{C}_0} \right)^{\alpha} \tag{12}$$

where $\mathcal{C}(\mathcal{D})$ represents the complexity of domain $\mathcal{D}$, and $\alpha$ denotes domain-specific scaling exponents that can be empirically determined.

This mathematical foundation transforms agency modeling from categorical classification to quantitative optimization, enabling systematic design of intelligent systems and principled comparison of natural and artificial intelligence across scales and domains. The resulting framework provides both theoretical insights into the nature of intelligence and practical tools for engineering more sophisticated artificial agents.