Chaotic Itinerancy in Collective Behavior Emerging from Active Inference: A Multi-Agent Model of Trust and Empowerment Dynamics in Theatre Workshops

Shoko Miyano * and Takashi Shiono

Posted Date: 12 March 2026

doi: 10.20944/preprints202603.0933.v1

Keywords: active inference; chaotic itinerancy; multi-agent systems; trust dynamics; free energy principle; theatre workshops

 doi:10.20944/preprints202603.0933.v1

Article

Chaotic Itinerancy in Collective Behavior Emerging from Active Inference: A Multi-Agent Model of Trust and Empowerment Dynamics in Theatre Workshops

Shoko Miyano [1,*] and Takashi Shiono [2]

[1] J. F. Oberlin University, Tokyo, Japan
[2] University of Tokyo, Tokyo, Japan
* Correspondence: miyano_s@obirin.ac.jp

Abstract

Chaotic itinerancy—irregular switching among metastable collective states—provides a dynamical substrate for flexible social coordination, yet its mechanistic origin in multi-agent systems remains unclear. We present a multi-agent Active Inference model in which chaotic itinerancy emerges from Expected Free Energy minimization without outcome-level social priors. Agents select actions to minimize Expected Free Energy while updating preferences through a precision-gated learning mechanism modulated by interpersonal trust. Hill-function nonlinearity in state transitions creates bistable "affordance landscapes" that gate behavioral mode switching. Simulations with small number of agents on an Erdős–Rényi trust network reveal spontaneous alternation among multiple metastable behavioral clusters, heavy-tailed dwell-time distributions, and sign-changing finite-time Lyapunov exponents—three hallmarks of chaotic itinerancy. Crucially, replacing Hill-function dynamics with linear transitions reduces the chaotic-itinerancy detection rate from 80% to 20%, demonstrating that nonlinear affordance structure is necessary for generating metastable switching. We further show that agents with simplified internal models of the world sustain richer itinerant dynamics as a group than "perfect-foresight" agents, suggesting that bounded rationality may be functionally advantageous for maintaining behavioral flexibility. These results establish active inference as a principled framework for modeling chaotic itinerancy in social systems and offer a computational account of trust-mediated collective transitions observed in theatre workshops and group dynamics.

Keywords: active inference; chaotic itinerancy; multi-agent systems; trust dynamics; free energy principle; theatre workshops

---

Lead Paragraph: Groups often shift abruptly between qualitatively different collective states—from hesitant silence to energetic participation, or from rigid conformity to creative exploration. Understanding when and why such transitions occur is central to the study of group dynamics and social interaction. Here we show that a specific form of structured irregularity, called chaotic itinerancy, can emerge spontaneously in a computational model where agents make decisions by minimizing uncertainty about their future. The key ingredient is a nonlinear threshold effect: certain behaviors become viable only when a sufficient level of mutual trust is established among group members, creating multiple "attractor ruins" that the system visits transiently before moving on. Our model captures dynamics observed in theatre workshops, where groups spontaneously alternate between phases of risk-taking and consolidation, and provides a mechanistic explanation for how interpersonal trust shapes the timing and character of collective mode switching.

1. Introduction

Understanding how abrupt collective shifts emerge from individual decision-making constitutes a central problem in complex social systems research. Many social phenomena exhibit nonlinear

 doi:10.20944/preprints202603.0933.v1

transitions whereby gradual individual-level changes yield sudden macroscopic shifts, including cascades and coordination breakdowns [1,2].

The Free Energy Principle (FEP) describes biological agents as minimizing variational free energy [3]. Within this framework, Active Inference casts action selection as Expected Free Energy (EFE) minimization, integrating goal-directed behavior and epistemic exploration [4].

Extending Active Inference to social systems raises a familiar concern. Some models introduce explicitly social or normative priors that directly favor particular collective outcomes ("social priors"), which can risk circularity if the target regularities are effectively encoded in prior preferences [5, 6]. We therefore ask whether chaotic itinerancy—structured switching among multiple metastable collective modes—can arise under EFE minimization when agents have minimally social preferences but structured generative models.

In theatre workshops and experimental actor training practices, the relevant shift is not limited to changes in average participation, but involves a qualitative transformation in how participants take relational risks and experience agency. Theatre theory and experimental practice have described such transformations in terms of the removal of habitual defenses (Grotowski) and entry into a precarious "empty space" in which action is no longer fully guided by established roles or expectations (Brook) [7,8]. Accordingly, we target chaotic itinerancy at the collective level—where groups alternate among multiple metastable modes before settling—and an individual-level learning mechanism that operationalizes "threshold crossing" through adaptive preferences.

Methodological stance: avoiding outcome-level social priors.

Active Inference social models sometimes place normative structure at the outcome or policy level. For example, Constant et al. formalize social conformity using *deontic value* and *deontic cues*, where observations can directly endow policies with normative salience (i.e., an observation-conditioned prior over "what one should do") [5]. Relatedly, models of cooperative communication may posit an *adaptive prior belief* that agents' mental states are aligned, and cast communicative acts as evidence gathering for that prior [6].

In this paper, we adopt a constraint that makes the explanatory target explicit. *We avoid* outcome-level priors that directly prescribe collective patterns (e.g., explicit conformity utilities or preferred participation outcomes). *Instead*, we place structure in latent-state dynamics and learning conditions (nonlinear collective effects and trust-gated preference plasticity) while keeping preferences over broadly non-normative state variables (trust, empowerment, stamina).

1.1. Why Theatre Workshops as a Model System for Social Interaction

Theatre performances serve as experimentally controllable microcosms of social coordination. This perspective is grounded in theatre theory and laboratory-based experimental practices that treat the rehearsal space as a laboratory for investigating human interaction.

According to Brook's concept of the "empty space," theatre can arise from the minimal configuration of a performer and a spectator. In this most condensed form, theatre makes visible the essential dynamics of human interaction—the continuous chain of action and reaction—by isolating them from the distractions of everyday social contexts[8].

Extending this experimental lineage, Grotowski's "Laboratory Theatre" explicitly conceptualized performance as research into the actor-spectator relationship, emphasizing a "via negativa" approach: not the accumulation of expressive techniques, but the systematic removal of habitual psychological and physical defenses that constrain action. By stripping away social masks and conditioned responses, performance becomes a site for exposing underlying impulses and relational dynamics[7].

Boal further extended this trajectory by dissolving the boundary between actor and audience, framing theatre as a space in which participants can actively explore alternative modes of interaction[9].

Taken together, these theatre theories and experimental practices converge on a shared view of performance as an experimentally constrained setting for probing the fundamental dynamics of social interaction, where relational patterns can be isolated, transformed, and explored. Crucially, this

 doi:10.20944/preprints202603.0933.v1

experimental orientation shifts attention away from representational display toward the conditions under which interaction itself is generated.

Building on these perspectives, this paper focuses on theatre workshops rather than traditional theatrical performances. While conventional theatre foregrounds the interaction between performers and spectators, the workshops considered here deliberately minimize the element of being observed.

To clarify what is meant by a theatre workshop in this context, we briefly outline a typical structure below. Theatre workshops vary widely in form, and many are designed not for trained actors but for participants with little or no prior experience in theatre. A typical workshop for such participants often begins with introductory activities aimed at building interpersonal relationships, followed by stretching and movement-based exercises, and then incorporates a series of expressive activities using improvisational acting that foreground mutual responsiveness. This shift redirects participants' attention away from presentation and evaluation, and toward the ongoing dynamics of action and response among those involved. Collectively, these features highlight how theatre workshops deliberately isolate and intensify coordination under shared uncertainty, providing a structured setting in which collective phenomena can be examined.

In this sense, the theatre workshop functions as a structured experimental environment, allowing collective patterns of interaction to emerge through embodied engagement rather than scripted representation. These considerations motivate our use of theatre workshops as a concrete experimental system for formal modeling and serve as constraints on model design.

In particular, we operationalize "crossing a threshold" as an internally learned change in preference parameters: empowerment overshoots, when experienced under sufficient inferred trust, are consolidated through precision-weighted hierarchical Bayesian learning and persist after the perturbation ends. This provides a minimal account of how a group can shift from inhibited participation to a stable mode of expressive agency without prescribing the outcome.

1.2. Why Chaotic Itinerancy Is Essential in Modeling Theatre Workshops

Theatre workshops rarely converge to a single stable collective mode. Instead, groups often alternate among qualitatively different regimes—inhibited, exploratory, playful, confrontational—with irregular transitions that depend on interaction history. This pattern is well described by *chaotic itinerancy* (CI): trajectories that dwell near multiple quasi-stable "attractor ruins" and switch among them in a structured but unpredictable manner [10–12]. Notably, a connection between active inference and chaos control was established early in the FEP literature: prediction-error suppression under strong priors can itself generate and stabilize chaotic trajectories [13].

CI provides an interpretable vocabulary for small-group processes. Rather than assuming that theatre workshops must converge to either "success" or "failure," CI emphasizes the transient, exploratory nature of collective dynamics: multiple metastable modes coexist, and facilitation may shape which modes are visited and in what sequence. This perspective aligns with theatre theory and experimental practice that value the process of exploration over fixed outcomes [7,8].

Contributions.

This paper makes two primary contributions:

- We demonstrate **Emergent Chaotic Itinerancy** in a multi-agent POMDP under Active Inference. This result successfully reproduces the dynamic process of theatre workshops—where groups alternate between periods of high energy and quiescence—providing a novel computational approach to theatre workshop analysis that has not been seen in previous studies.
- We introduce **Precision-Gated Preference Learning** as a natural implementation of preference inference within Active Inference, enabling the agent's preference model to be discontinuously updated by large surprise events via a hierarchical Bayesian mechanism.

To realize these contributions, the model incorporates the following distinctive features:

 doi:10.20944/preprints202603.0933.v1

- **Hill Function Nonlinearity**: We embed cooperative threshold effects via the Hill function in state transition dynamics, which enables chaotic itinerancy to emerge. Ablation experiments confirm that removing this nonlinearity (reducing to a linear AR(1) process) eliminates chaotic itinerancy entirely—demonstrating that the Hill function is the essential mechanism generating complex collective dynamics.
- **Interpersonal Trust Network**: We extend mean-field dynamics to a sparse Erdős–Rényi graph where dyadic trust evolves based on action synchrony and local state variables are spatially averaged, enabling heterogeneous social perception.
- **Robust Parameter Regime**: Systematic scans confirm that the qualitative behavior is robust across a wide range of parameters.

We study a stylized setting inspired by theatre workshops, where participants decide whether to express themselves creatively under partial observability and resource constraints. The same formal structure can be applied to other social systems in which a shared latent context (here: trust) interacts with individual agency and fatigue. The remainder of the paper specifies the POMDP model and EFE-based policy, reports numerical evidence for chaotic itinerancy, and discusses implications and empirical extensions.

2. Related Work on Collective Phase Transitions in Small-Group Systems

2.1. Sociophysics and Ising-Type Perspectives on Collective Transitions

A large body of work in sociophysics and statistical-physics-inspired social modeling has demonstrated that sharp collective transitions, multistability, and hysteresis can arise from simple local interaction rules, including Ising-like binary-choice dynamics and threshold/cascade models [1,2]. These models provide an important baseline vocabulary for understanding phase-transition-like phenomena in social systems. However, they are typically formulated either as population-level stochastic dynamics with externally specified update rules (e.g., Glauber-type flips), or as utility-like response functions, rather than as agents performing principled inference and planning under a generative model (i.e., explicit modeling of agency).

2.2. Small-Group Dynamical Models with Metastable Collective States

Closer to the scale of workshops and therapeutic groups, quantitative small-group models have explicitly investigated the emergence and alternation of metastable collective states. Notably, Lauro Grotto et al. proposed an analytic dynamical-systems model of interacting agents (with simplified emotional/cognitive variables) and studied when metastable group-level states emerge, how these states differ structurally, and how system size affects the dynamics [14]. This line of work is especially relevant because it targets the "small-N" regime (comparable to workshop groups) and treats the group as a setting in which qualitatively distinct collective modes can persist and switch.

2.3. Psychotherapy and Therapeutic Change as Nonlinear Dynamics and Critical Transitions

Independent of theater, psychotherapy research has developed a mature complex-systems tradition in which therapeutic change is treated as nonlinear, sometimes discontinuous, and potentially characterized by critical instabilities and phase-transition-like shifts. A representative mathematical approach is the dynamical model of psychotherapy proposed by Liebovitch et al. [15]. Empirically and conceptually, the synergetic/complex-systems program associated with Schiepek and collaborators emphasizes self-organization, critical fluctuations, and abrupt transitions in therapy process time series [16–18]. These studies provide precedent for treating "intervention + interaction history" as capable of moving a social-therapeutic system across qualitative regimes, i.e., path dependence.

While these psychotherapy models are not specifically about theatre workshops, their methodological stance is directly aligned with the present aim: explaining multistability and history dependence as emergent properties of coupled human processes.

 doi:10.20944/preprints202603.0933.v1

2.4. Computational and Agent-Based Perspectives on Improvisation and Performance

Research on improvisational performance has also employed computational and agent-based perspectives to formalize how collective structure can arise from moment-to-moment interaction. For example, empirical and computational work in the creativity-and-cognition community has analyzed cognition and interaction in theatrical improvisation as a domain of situated, co-constructed action [19]. Such approaches motivate the use of performance and workshop settings as experimentally controllable microcosms of social coordination, while also highlighting the relative scarcity of computable models that simultaneously capture (i) latent-field-like variables (e.g., trust/safety), (ii) resource constraints (e.g., fatigue/stamina), and (iii) irreversible learning effects at the preference level.

2.5. Theatre, Predictive Processing, and Active Inference: Scope of Existing Work

Work connecting theatre and performance to predictive processing and the Free Energy Principle has begun to articulate how engagement, affordances, and spectatorship can be interpreted through prediction-based frameworks. For example, Murphy discusses theatrical experience in terms of predictive engagement and imaginative play within an affordance landscape [20]. These contributions provide useful conceptual bridges, but they do not typically instantiate a multi-agent generative model that can be simulated and evaluated against dynamical signatures (e.g., chaotic itinerancy) in workshop-like group interaction.

In a nearby methodological direction, Active Inference has been used to formalize social structure in terms of shared scripts and action sequences. Albarracin et al. propose a variational formalization of scripts within Active Inference [21], offering a natural point of contact with role- and script-like regularities in performance settings. However, within the scope of our literature search, we did not identify peer-reviewed work that formulates theatre workshops (or closely related drama-therapeutic group dynamics) as an explicit computational state-space/POMDP model under Active Inference and then systematically evaluates emergent collective dynamics via simulation.

2.6. Positioning of the Present Work

Taken together, prior work establishes (a) that collective phase-transition-like behavior is common in abstract social models and (b) that small-group and psychotherapy contexts can exhibit metastable regimes and discontinuous shifts. It also suggests (c) that improvisational performance is amenable to computational formalization. The present study contributes a complementary perspective by formulating workshop dynamics as a multi-agent POMDP under Active Inference: chaotic itinerancy—structured switching among multiple metastable collective modes—arises under EFE minimization with structured state transitions and trust-gated learning, providing a principled account of how groups alternate among transient modes without prescribing collective outcomes.

3. Theoretical Framework: Active Inference and Expected Free Energy

3.1. Free Energy Principle: Basic Structure

The Free Energy Principle (FEP) describes the behavior of biological agents as "surprise minimization" [3]. Da Costa et al. [4] formalized this principle as a normative theory of action selection, demonstrating that **Active Inference** can be rigorously derived under physically plausible assumptions.

Their core insight is that under the **precise agent** assumption—macroscopic biological agents respond deterministically to their environment—agent behavior can be described as Expected Free Energy (EFE) minimization.

 doi:10.20944/preprints202603.0933.v1

Definition 1 (Precise Agent [4]). *A precise agent is one for which the external state s determines both the observation o and the action a:*

$$P(o \mid s, e, h_{\leq t}) = \delta_{f(s)}(o) \tag{1}$$

$$P(e \mid s, o, h_{\leq t}) = \delta_{g(s)}(e) \tag{2}$$

where δ denotes the Dirac delta function and $h_{\leq t}$ is the agent's history.

3.2. Expected Free Energy: Definition and Decomposition

According to Da Costa et al. [4], an Active Inference agent is completely characterized by two models:

- **Predictive model** $P(s, o \mid a)$: Joint distribution of external states s and observations o given action a.
- **Preference model** $P(s, o)$: Distribution over preferred external states and observations.

Action selection is governed by the Expected Free Energy:

$$-\log P(a|h) = \mathbb{E}_{P(s,o|a,h)}[\log P(s|a,h) - \log P(s,o|h)] \tag{3}$$

The EFE admits two equivalent decompositions:

Risk–Ambiguity Decomposition:

$$-\log P(a|h) = \underbrace{D_{\mathrm{KL}}[P(s|a,h) \| P(s|h)]}_{\text{Risk}} + \underbrace{\mathbb{E}_{P(s|a,h)}[H[P(o|s,h)]]}_{\text{Ambiguity}} \tag{4}$$

- **Risk**: KL divergence between predicted and preferred external states. Minimizing this reverse KL divergence induces **mode-matching behavior**, providing the mathematical foundation for **risk-averse decision making**.
- **Ambiguity**: Expected entropy of future observations. High ambiguity indicates uncertainty about what will occur. Ambiguity minimization drives exploration toward states where observations clearly reveal external states (the **streetlight effect**).

Extrinsic–Intrinsic Value Decomposition:

$$-\log P(a|h) \geq \underbrace{-\mathbb{E}_{P(o|a,h)}[\log P(o|h)]}_{\text{Extrinsic Value}} \underbrace{-\mathbb{E}_{P(o|a,h)}[D_{\mathrm{KL}}[P(s|o,a,h) \| P(s|a,h)]]}_{\text{Intrinsic Value}} \tag{5}$$

- **Extrinsic Value**: Log-likelihood of preferred observations. It corresponds to **reward or utility maximization**, the foundation of reinforcement learning and Bayesian decision theory.
- **Intrinsic Value**: Information gain from action (mutual information). Provides the mathematical basis for **curiosity-driven exploration and active learning**, corresponding to Bayesian experimental design.

3.3. Three Advantages of Active Inference

Da Costa et al. [4] identify three key advantages of Active Inference:

(a) **Principled resolution of exploration-exploitation dilemma**: Information gain (intrinsic value) is naturally incorporated into action selection, without requiring separate exploration bonuses.

(b) **Explainable action simulation under generative world models**: Actions are explained as consequences of beliefs about the world, enabling interpretable agent behavior.

(c) **Universality**: Any RL algorithm can be re-described as Active Inference, suggesting AI as a unifying framework for sequential decision making.

 doi:10.20944/preprints202603.0933.v1

3.4. Position of This Study

This study applies the theoretical framework to **multi-agent collective behavior**, specifically demonstrating that **nonlinearity in the predictive model** and **coupling between state variables** can generate chaotic itinerancy—structured switching among multiple metastable collective modes.

One modeling approach for realizing complex collective behavior is to introduce explicit **Social Priors** (social prior distributions) directly into action selection. Here, by incorporating Hill function cooperative effects and inter-variable coupling into **the state transition dynamics themselves**, we show that chaotic itinerancy emerges from pure EFE minimization, with trajectories dwelling near multiple quasi-stable "attractor ruins" and switching irregularly among them.

3.5. Hill Function Nonlinearity

To model cooperative effects that create phase transitions, we employ the Hill function [22]:

$$\mathcal{H}_n(e;K) = \frac{e^n}{e^n + K^n} \tag{6}$$

where n is the Hill coefficient (cooperativity index) and K is the half-saturation constant (critical threshold).

Definition 2 (Hill Function)**.** *The Hill function $\mathcal{H}_n(e;K)$ has the following properties:*

- $\mathcal{H}_n(0;K) = 0, \quad \mathcal{H}_n(1;K) = \dfrac{1}{1+K^n}$
- $\mathcal{H}_n(K;K) = 0.5$ *(half-maximum at $e = K$)*
- *Slope at $e = K$:* $\left.\dfrac{d\mathcal{H}_n}{de}\right|_{e=K} = \dfrac{n}{4K}$
- *Converges to a step function as $n \to \infty$*
- *$n \geq 4$ is a necessary condition for bistability (a minimal form of multistability) in this model*

Originally developed to model cooperative binding in biochemistry, the Hill function has properties ideal for modeling collective effects:

- Sigmoidal response with threshold behavior at $e = K$
- Sharp transitions for $n \geq 4$, enabling bistability (and, more generally, multistable regimes)
- Continuous approximation of threshold dynamics

Affordance interpretation: Crucially, the Hill function specifies *environmental constraints on action feasibility*—an affordance structure—rather than prescribing preferred outcomes. It encodes the physical and psychosocial conditions under which sustained collective expression becomes viable: when group expression exceeds threshold K, the environment *affords* reduced effort costs and enhanced trust gains. This is analogous to how terrain affords walking or obstacles constrain movement—the structure shapes what actions are sustainable, not what actions are desirable. Thus, embedding Hill nonlinearity in state transitions differs fundamentally from introducing outcome-level social priors that directly encode conformity or coordination as preferred observations.

4. Model Specification

This section formulates a discrete-time multi-agent model abstracting a theatre workshop. A distinctive feature of this model is that it incorporates Hill function nonlinearity and inter-variable coupling into **the predictive model itself**, so that agents possess this dynamics as an internal model and reflect it in action selection through the Risk term of EFE.

All simulations were implemented in Python using standard numerical libraries.

 doi:10.20944/preprints202603.0933.v1

4.1. POMDP Structure

The model is a Partially Observable Markov Decision Process (POMDP) with the following structure:

State Space Classification

For agent i, states are classified into three types:

- **External states (partially observable)**: S_t (collective trust) and $\{W_{ij,t}\}_{j \in \mathcal{N}(i)}$ (interpersonal trust). S_t is a "field state" shared by all agents, while $W_{ij,t}$ represents dyadic relational strength. Both are latent and must be inferred from observations.
- **Internal states (fully observable)**: (u_t^i, H_t^i)—Empowerment u_t^i and stamina H_t^i are the agent's own states, which the agent can fully know.

POMDP Components

Table 1. POMDP components of the workshop model.

Component	Symbol	Description
External states	$S_t, \{W_{ij,t}\}$	Collective trust, Interpersonal trust (partially observable)
Internal states	(u_t^i, H_t^i)	Empowerment, stamina (fully observable)
Observation	o_t^i	Perceived others' expression rate (noisy)
Action	$a_t^i \in \{0, 1, 2\}$	Rest / Chat & Exercise / Express
History	$h_{<t}^i$	$(o_{<t}^i, a_{<t}^i, u_{<t}^i, H_{<t}^i)$

Observability of State Variables

- S_t (collective trust) and $W_{ij,t}$ (interpersonal trust) are latent variables representing the "atmosphere" and "relationship quality," respectively. These cannot be directly measured and must be inferred from social signals (others' expression and chat & exercise responses).
- u_t^i (empowerment) is a subjective feeling of "how empowered one is," which is fully graspable through self-introspection.[1]
- H_t^i (stamina) is one's own bodily state, which is fully self-aware.

4.2. State Variables

This subsection introduces the set of latent and internal variables required to express the workshop dynamics as a tractable multi-agent POMDP on a network. We distinguish between a shared, partially observable field state S_t (collective trust), dyadic relational states $W_{ij,t}$ (interpersonal trust), and agent-specific internal states (u_t^i, H_t^i) (empowerment and stamina). Inter-agent coupling is structured by a sparse network, where agents perceive a local trust-weighted average of others' behavior. This design choice makes explicit what agents must infer (collective and interpersonal trust) versus what they can access directly (their own empowerment and stamina), which is central for the epistemic term in Expected Free Energy.

Indices

- i: Agent (participant) index $(i = 1, \ldots, N)$
- t: Discrete time step $(t = 0, 1, 2, \ldots, T)$
- τ: Relative time within planning horizon $(\tau = 1, \ldots, N_{\text{horizon}})$

[1] We use "empowerment" in its psychological sense—a felt sense of competence, self-determination, and capacity to act effectively [23,24]—rather than in the information-theoretic sense of channel capacity between actions and future states [25]. While both concepts share the intuition that agents seek states enabling effective action, the information-theoretic definition quantifies potential control in bits, whereas our usage captures the subjective experience of self-efficacy that influences action selection through preferences.

 doi:10.20944/preprints202603.0933.v1

State Variables

- $S_t \in \mathbb{R}$: Collective trust (Gaussian latent variable)
- $u_t^i \in \mathbb{R}$: Agent i's empowerment (Gaussian latent variable)
- $H_t^i \in [0, H_{\max}]$: Agent i's stamina
- $W_{ij,t} \in \mathbb{R}$: Interpersonal trust between agents i and j (network edge weight)

Network Structure

Agents are embedded in a sparse undirected network $G = (V, E)$ with adjacency matrix $A_{ij} \in \{0, 1\}$, where $A_{ij} = A_{ji}$. The network is generated as an Erdős–Rényi random graph with connection probability $p = k_{\text{avg}}/(N-1)$, where k_{avg} is the target average degree (default: $k_{\text{avg}} = 4$). Each edge $(i, j) \in E$ carries a dynamic interpersonal trust weight $W_{ij,t} \in \mathbb{R}$ that evolves based on action synchrony (see Section 4.3.0.5). The neighborhood of agent i is denoted $\mathcal{N}(i) = \{j : A_{ij} = 1\}$.

This network structure serves two purposes:

1. **Avoiding complete mean-field approximation**: Rather than assuming all agents observe the same global expression rate $\bar{e}_t$, each agent perceives a *local* expression rate computed over its neighborhood.
2. **Maintaining computational tractability**: The sparse network topology ($k_{\text{avg}} \ll N$) and local averaging keep inference tractable while introducing heterogeneous social perception.

Gaussian Latent Variable Design

A key design principle is that the primary state variables S_t and u_t^i are modeled as **Gaussian latent variables** without artificial clipping to $[0, 1]$. This ensures theoretical consistency with the Gaussian assumptions underlying belief updating (Extended Kalman Filter) and information gain computation. Physical effects that require bounded values are obtained through sigmoid transformations:

$$\text{Bounded effect} = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{7}$$

Interpretation of Unbounded States:

- **Negative S (distrust)**: Values $S < 0$ represent collective distrust or psychological unsafety. The sigmoid transformation ensures that negative trust does not produce paradoxical amplification effects.
- **Negative u (disempowerment)**: Values $u < 0$ represent a sense of powerlessness or learned helplessness.
- **Large positive values**: Values significantly above 1 represent strong trust or high empowerment, with diminishing marginal physical effects due to sigmoid saturation.

This design choice keeps the model's state dynamics and belief updating aligned with the Gaussian assumptions used in inference (Extended Kalman Filter) and in the information gain computation. The information gain expression $I = \frac{1}{2}\log(1 + H^2 v/R)$ is exact under linear–Gaussian assumptions. Accordingly, we keep S_t and u_t^i unconstrained and apply sigmoid transformations only when mapping latent variables to bounded interaction terms, maintaining internal consistency of the computations.

Action Variable

- $a_t^i \in \{0, 1, 2\}$: Agent i's action (0: Rest, 1: Chat & Exercise, 2: Express)
- $e_t^i \in \{0, 1\}$: Expression indicator, $e_t^i = \mathbf{1}[a_t^i = 2]$

The three-valued action space distinguishes between:

- **Rest** ($a = 0$): Passive observation with stamina recovery.
- **Chat & Exercise** ($a = 1$): Low-cost social interaction that provides *information gain* about interpersonal trust W_{ij} (via IG_W) and also updates W_{ij} through action synchrony.

 doi:10.20944/preprints202603.0933.v1

- **Express** ($a = 2$): High-commitment creative expression that consumes stamina, builds collective trust S via the Hill-function mechanism, and updates interpersonal trust W_{ij} through action synchrony—but does *not* provide information gain IG_W (see Supplementary Material (Sec. S-X) for sensitivity analysis of this design choice).

Collective Statistics

The model supports two modes of collective coupling:
(a) Global statistics (used for collective trust S dynamics):

$$\bar{e}_t = \frac{1}{N}\sum_{j=1}^{N} e_t^j \qquad \text{(overall expression rate)} \tag{8}$$

$$\bar{c}_t = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}[a_t^j = 1] \qquad \text{(overall chat rate)} \tag{9}$$

(b) Local statistics (used for empowerment and stamina dynamics):

$$\bar{e}_t^{(i)} = \frac{\sum_{j\in\mathcal{N}(i)} \sigma(W_{ij,t}) \cdot e_t^j}{\sum_{j\in\mathcal{N}(i)} \sigma(W_{ij,t}) + \varepsilon} \tag{10}$$

where $\mathcal{N}(i)$ is agent i's neighborhood, $\sigma(\cdot)$ is the sigmoid function, and $\varepsilon > 0$ is a small constant for numerical stability.

The local expression rate $\bar{e}_t^{(i)}$ is a **trust-weighted average** over the agent's neighborhood: neighbors with higher interpersonal trust W_{ij} contribute more to the perceived social context. This replaces the mean-field statistic $\bar{e}_t^{-i}$ used in earlier formulations, enabling heterogeneous social perception while maintaining computational tractability.

Information Asymmetry: Global vs. Local Coupling

The model employs an intentional **asymmetry in information scope** that reflects the physical structure of theatre workshop settings:

- **Global coupling for collective trust** S: Expressive acts (Express) are publicly visible performances that all participants can observe simultaneously—akin to stage performances visible to the entire room. Accordingly, the collective trust S is updated based on the *global* expression rate $\bar{e}_t$ (Eq. 11).
- **Local coupling for empowerment** u^i **and stamina** H^i: In contrast, the psychological impact of others' expression on one's own sense of empowerment and fatigue depends on *whom one is attending to*—typically nearby participants with whom one has established interpersonal trust. Hence, empowerment (Eq. 12) and stamina dynamics use the *local* trust-weighted expression rate $\bar{e}_t^{(i)}$.
- **Local coupling for interpersonal trust** W_{ij}: Dyadic trust evolves based on pairwise action synchrony between connected agents (Eq. 15).

This two-layer structure—global media (publicly visible expression affecting shared trust climate) combined with local communication (interpersonal relationships affecting individual empowerment)—generalizes beyond theatre workshops. Analogous patterns appear in organizational settings (company-wide announcements vs. team-level interactions), social media (viral content vs. direct messaging), and community dynamics (public gatherings vs. neighborhood conversations).

Belief Variables (Perception)

In Active Inference, agents minimize variational free energy to approximate the posterior distribution over hidden states given observations [3,4]. Since collective trust S_t and interpersonal trust $W_{ij,t}$ are not directly observable, agent i maintains a probabilistic belief $Q(s)$ over them. We employ a Gaussian approximation for these beliefs, parameterized by their sufficient statistics (mean and

 doi:10.20944/preprints202603.0933.v1

variance). This belief state constitutes the agent's "perception" of the social environment and serves as the basis for calculating the epistemic value (information gain) of future actions.

- $m_S^i, v_S^i \in \mathbb{R} \times \mathbb{R}_{>0}$: Mean and variance of belief about collective trust S
- $m_{W_{ij}}^i, v_{W_{ij}}^i \in \mathbb{R} \times \mathbb{R}_{>0}$: Mean and variance of belief about interpersonal trust W_{ij}

Preference Variables (Goal Directedness)

Active Inference unifies perception and action under the single imperative of minimizing free energy. Goal-directed behavior arises from the agent's *prior preferences* over states, denoted as $P(s)$. Action selection minimizes Expected Free Energy (EFE), which includes a "Risk" term defined as the KL divergence between the predicted state distribution and the preferred distribution $P(s)$ [4]. In this model, agents possess preferred setpoints for trust, empowerment, and stamina. Crucially, unlike fixed reward functions in standard RL, these preference parameters (specifically μ_U^i) can themselves be updated through learning (see Section 4.7).

- $\mu_S^i \in \mathbb{R}$: Preferred collective trust level
- $\mu_U^i \in \mathbb{R}$: Preferred empowerment level (adaptively learned)
- $\mu_H^i \in [0, H_{\max}]$: Preferred stamina level

4.3. State Transition Dynamics

We specify a discrete-time generative process in which (i) collective trust evolves as a leaky (AR(1)) process with a nonlinear, threshold-like social amplification term, (ii) empowerment accumulates through self- and other-driven gains modulated by trust, and (iii) stamina implements a simple resource constraint with recovery and cooperative cost reduction.

Collective Trust:

$$S_{t+1} = a_S S_t + b_S + c_S^{\text{Hill}} \cdot \mathcal{H}_n(\bar{e}_t; K) + c_S^{\text{chat}} \cdot \bar{c}_t + \omega_t \tag{11}$$

where $\omega_t \sim \mathcal{N}(0, \sigma_S^2)$ is process noise, $\bar{e}_t$ is the global expression rate (Eq. 10), and $\bar{c}_t$ is the global chat rate.

Here, $a_S S_t + b_S$ captures gradual decay toward a baseline, the Hill term $c_S^{\text{Hill}} \mathcal{H}_n(\bar{e}_t; K)$ implements a sharp increase in trust once expression becomes sufficiently common (critical threshold K), and the linear term $c_S^{\text{chat}} \bar{c}_t$ provides a modest contribution from Chat & Exercise activity.

Parameters:

- a_S: Decay coefficient ($0 < a_S < 1$)
- b_S: Baseline constant
- c_S^{Hill}: Hill function effect strength (Express)
- c_S^{chat}: Linear effect strength (Chat & Exercise)

The Hill function term captures the nonlinear amplification of trust when group expression exceeds the critical threshold K. Conceptually, this corresponds to the idea that a psychologically safe climate can emerge nonlinearly once sufficient interpersonal risk-taking (here, expressions) becomes common [26,27].

Empowerment (with Trust–Empowerment Coupling):

$$\begin{aligned} u_{t+1}^i = \alpha_U u_t^i + \eta_{\text{self}} e_t^i + \eta_{\text{other}} \bar{e}_t^{(i)} \left(1 + \gamma_{\text{coop}} \cdot \sigma(u_t^i)\right) \\ + \underbrace{\kappa_{S \to u} \cdot \sigma(S_t) \cdot \bar{e}_t^{(i)}}_{\text{Trust–Empowerment Coupling}} + \xi_t^i \end{aligned} \tag{12}$$

 doi:10.20944/preprints202603.0933.v1

where $\xi_t^i \sim \mathcal{N}(0, \sigma_U^2)$ is process noise, $\sigma(\cdot)$ is the sigmoid function (Eq. 7), and $\bar{e}_t^{(i)}$ is the local expression rate (Eq. 10).

The cooperative factor uses sigmoid-transformed empowerment $\sigma(u_t^i)$ to prevent runaway positive feedback. The trust–empowerment coupling term makes vicarious gains systematically larger when inferred trust is high. Both sigmoid transformations ensure bounded contributions regardless of the magnitude of the underlying Gaussian latent variables.

Parameters:

- α_U: Decay coefficient ($0 < \alpha_U < 1$)
- η_{self}: Empowerment increase from self-expression
- η_{other}: Empowerment increase from others' expression
- γ_{coop}: Cooperative amplification strength
- $\kappa_{S \to u}$: Trust–empowerment coupling coefficient

Theoretical Significance of Trust–Empowerment Coupling:

The underlined term $\kappa_{S \to u} \cdot \sigma(S_t) \cdot \bar{e}_t^{(i)}$ represents that when the field trust S is high, the positive effect of neighbors' expression on one's own empowerment is amplified.

Treatment of Negative Trust: The trust variable S_t is modeled as a Gaussian latent variable (Eq. 11) and can take any real value. Physical effects that depend on trust use sigmoid-transformed values, ensuring bounded contributions:

$$S_{\text{eff}} = \sigma(S_t) = \frac{1}{1 + e^{-S_t}} \tag{13}$$

This design has several advantages:

1. **Theoretical consistency**: The Kalman filter belief updating and information gain formulas remain exact for the underlying Gaussian process.
2. **Asymmetric interpretation**: Negative trust (distrust) produces near-zero effective coupling, while positive trust produces positive coupling. This captures the psychological observation that distrust does not symmetrically reverse trust effects—rather, it attenuates or nullifies them.
3. **Smooth dynamics**: The sigmoid transformation is differentiable everywhere, avoiding discontinuities that could destabilize the dynamics.

This coupling term models several interrelated psychological phenomena. In trusted environments, others' expressions more readily empower the observer through empathic resonance—witnessing a breakthrough can feel as if it were one's own—and through imitation and contagion of expressive content [26,27]. Social learning is facilitated: others' successful experiences transfer more easily to one's own self-efficacy in safe contexts [24,28], and supportive environments promote internalization and autonomous motivation [29]. The resulting interaction between S and u creates positive feedback that stabilizes multistability (here: bistability).

Stamina (with Cooperative Cost Reduction):

$$H_{t+1}^i = \begin{cases} \min(H_t^i + H_{\text{rec}}, H_{\max}) & \text{if } a_t^i = 0 \text{ (rest)} \\ H_t^i - c_{\text{chat}} & \text{if } a_t^i = 1 \text{ (chat \& exercise)} \\ \max\left(H_t^i - c_{\text{exp}} \cdot \left(1 - c_H^{\text{Hill}} \cdot \mathcal{H}_{n_H}(\bar{e}_t^{(i)}; K_H)\right), 0\right) & \text{if } a_t^i = 2 \text{ (express)} \end{cases} \tag{14}$$

This piecewise update implements recovery during rest, mild depletion during chat & exercise, and significant depletion during expression, with costs reduced when neighbors are also expressing (cooperative cost reduction via local expression rate $\bar{e}_t^{(i)}$). The min/max operators keep stamina within $[0, H_{\max}]$, representing a hard resource constraint.

Parameters:

 doi:10.20944/preprints202603.0933.v1

- $H_{\max}$: Maximum stamina
- H_{rec}: Stamina recovery during rest
- c_{chat}: Chat & Exercise cost (small)
- c_{exp}: Basic expression cost
- c_H^{Hill}: Cooperative cost reduction coefficient
- n_H: Hill coefficient (stamina)
- K_H: Half-saturation constant (stamina)

The factor $(1 - c_H^{\text{Hill}} \cdot \mathcal{H}_{n_H}(\bar{e}_t^{(i)}; K_H))$ implements **cooperative cost reduction**: expression costs less when neighbors are also expressing. This positive feedback mechanism is key to enabling multistability. This stylized assumption is consistent with classic findings that the presence of others can systematically modulate effort and performance (social facilitation) [30], and with psychophysical accounts that treat effort as a subjective, reportable quantity (ratings of perceived exertion) [31].

Sociological Justification for the Hill Function: While the Hill function is originally derived from biochemistry (cooperative oxygen binding to hemoglobin), we employ it here as a phenomenological ansatz for **threshold-dependent social reinforcement**. This aligns with sociological theories of "critical mass" and "complex contagion" [32], which posit that adoption of costly or risky behaviors (like expressive performance) requires social reinforcement from multiple sources exceeding a specific threshold, unlike simple information spreading. The Hill coefficient n controls the steepness of this threshold, effectively modeling how strictly the group enforces this critical mass requirement.

From an ecological perspective, the Hill function formalizes an **affordance structure**: the collective environment either affords or constrains sustained expression depending on whether participation exceeds the critical threshold. This framing clarifies that we are specifying *what the environment makes possible*—a physical and psychosocial constraint—rather than *what agents should prefer*. The collective patterns that emerge are not built into preferences but arise from agents navigating this affordance landscape under EFE minimization.

Physiological Motivation for Hill-Type Stamina Dynamics

We use a Hill-type nonlinearity as a compact way to represent cooperative, threshold-like changes in endurance costs. The Hill function is classically associated with cooperative oxygen binding and the resulting sigmoidal saturation curve [22], and endurance physiology highlights oxygen transport as a key determinant of fatigue resistance [33]. By analogy, when group expression exceeds a critical threshold, a "collective oxygen supply" effect emerges: mutual entrainment and synchronization among participants—well documented in sports science and rhythmic coordination [30]—reduces the subjective effort required for sustained expression. Empirically, perceived effort and fatigue exhibit nonlinear dependence on physiological and contextual factors [31]. In this model the Hill term is not a literal haemoglobin model; it is an abstract mechanism by which supportive collective conditions can reduce the effective cost of expression once participation exceeds a critical level.

Interpersonal Trust Dynamics

In addition to the collective trust S_t, agents maintain dyadic interpersonal trust weights $W_{ij,t}$ on network edges. These weights evolve based on **action synchrony**—whether paired agents take coordinated actions:

$$W_{ij,t+1} = a_W W_{ij,t} + b_W + c_W \cdot \phi_{\text{base}}(a_t^i, a_t^j) \cdot \sigma(S_t) + \nu_{ij,t} \tag{15}$$

where $\nu_{ij,t} \sim \mathcal{N}(0, \sigma_W^2)$ is process noise and $\phi_{\text{base}}(a^i, a^j)$ is the **synchrony effect function**:

Interpretation:

- **Express–Express synchrony** ($\phi = +1.0$): Joint creative expression strongly builds interpersonal trust.

 doi:10.20944/preprints202603.0933.v1

Table 2. Synchrony effect function $\phi_{\text{base}}(a^i, a^j)$.

ϕ_{base}	Rest (0)	Chat & Ex. (1)	Express (2)
Rest (0)	-0.5	-0.5	-1.0
Chat & Ex. (1)	-0.5	$+0.5$	-1.0
Express (2)	-1.0	-1.0	$+1.0$

- **Chat & Exercise synchrony** ($\phi = +0.5$): Mutual conversation and exercise moderately builds trust.
- **Asynchrony** ($\phi < 0$): Mismatched actions (one expresses while the other rests) erode trust.

The trust modulation $\sigma(S_t)$ ensures that interpersonal trust builds more readily in high collective trust environments. Since $W_{ij} = W_{ji}$ (undirected network), both agents experience the same trust update.

Key distinction: All actions (Rest, Chat & Exercise, Express) affect the *true dynamics* of W_{ij} through the synchrony matrix ϕ_{base}. However, only Chat & Exercise provides *information gain* about W_{ij} (see Eq. 54). This separation means that Express can build interpersonal trust through joint creative activity, but agents cannot directly observe this trust-building; they can only learn about interpersonal trust through Chat & Exercise.

Network Parameters:

Table 3. Network parameters and default values.

Symbol	Description	Default
k_{avg}	Average network degree	4
a_W	Interpersonal trust decay	0.90
b_W	Interpersonal trust bias	0.00
c_W	Synchrony effect strength	0.40
σ_W	Interpersonal trust noise	0.10

For belief updating over W_{ij}, agents use variational message passing (VMP) with Jaakkola bounds to handle the sigmoid nonlinearity in Eq. 10. Details are provided in Supplementary Material (Sec. S-XIII).

4.4. Predictive and Preference Models

Active Inference agents are completely characterized by two models:

(1) Predictive Model $P(s, o \mid a)$

Prediction of how external states and observations evolve under action $a \in \{0, 1, 2\}$. In this model:

$$P(S_{t+1} \mid S_t, \bar{e}_t, \bar{c}_t) = \mathcal{N}\big(a_S S_t + b_S + c_S^{\text{Hill}} \mathcal{H}_n(\bar{e}_t; K) + c_S^{\text{chat}} \bar{c}_t,\ \sigma_S^2\big) \tag{16}$$

$$P(o_t^i \mid S_t, a_t^i) = \mathcal{N}\Big(g(S_t),\ R(a_t^i)\Big) \tag{17}$$

where $g(\cdot)$ is the observation function (mapping to others' expression rate).

 doi:10.20944/preprints202603.0933.v1

(2) Preference Model $P(s, o)$

Distribution over preferred external states and observations. This model has preferences for all three state variables:

$$P_{\text{pref}}(S, u^i, H^i) \propto \exp\Big(-\frac{1}{2}\Big[w_S(S-\mu_S^i)^2 + w_U(u^i-\mu_U^i)^2 + w_H(H^i-\mu_H^i)^2\Big]\Big) \tag{18}$$

where:

- Preference for S: Preferred trust μ_S^i (belief-based since partially observable)
- Preference for u^i: Preferred empowerment μ_U^i (fully observable)
- Preference for H^i: Preferred stamina μ_H^i (fully observable)

4.5. Observation Model

Agents observe others' expression rate with noise:

$$o_t^i = \bar{e}_t^{(i)} + \epsilon_t^i, \quad \epsilon_t^i \sim \mathcal{N}(0, R(a_t^i)) \tag{19}$$

where $\bar{e}_t^{(i)}$ is the local expression rate (Eq. 10) and the observation noise depends on action:

$$R(a) = \begin{cases} R_{\text{rest}} & a = 0 \\ R_{\text{chat}} & a = 1 \\ R_{\text{express}} & a = 2 \end{cases} \quad (R_{\text{express}} < R_{\text{chat}} < R_{\text{rest}}) \tag{20}$$

This **active sensing** property means that more engaged actions (Express > Chat & Exercise > Rest) improve observation precision, linking action to information gain. Express provides the highest precision for observing collective trust S, while Chat & Exercise provides moderate precision but uniquely enables information gain about interpersonal trust W_{ij}.

4.6. Belief Update via Extended Kalman Filter

For the partially observable trust S, agents maintain Gaussian beliefs $Q(S_t) = \mathcal{N}(m_S^i, v_S^i)$. Upon receiving observation o_t^i, beliefs are updated via extended Kalman filter:

Prediction Step:

$$m_S^{i,\text{prior}} = a_S m_S^i + b_S + c_S^{\text{Hill}} \cdot \mathcal{H}_n(\hat{e}; K) \tag{21}$$

$$v_S^{i,\text{prior}} = a_S^2 v_S^i + \sigma_S^2 \tag{22}$$

Update Step:

$$K_t^i = \frac{H v_S^{i,\text{prior}}}{H^2 v_S^{i,\text{prior}} + R(a_t^i)} \tag{23}$$

$$m_S^{i,\text{post}} = m_S^{i,\text{prior}} + K_t^i \cdot (o_t^i - H m_S^{i,\text{prior}}) \tag{24}$$

$$v_S^{i,\text{post}} = (1 - K_t^i H) v_S^{i,\text{prior}} \tag{25}$$

where H is the observation matrix (gradient of the linearized observation model).

 doi:10.20944/preprints202603.0933.v1

4.7. Precision-Gated Preference Learning

A distinctive feature of our model is that preferences can update dynamically based on experience. This implements and extends **Preferential Inference** [4] to include adaptive preference parameter learning.

4.7.1. Theoretical Foundation: Hierarchical Bayesian Model

We formulate preference learning as variational inference over a hierarchical generative model. Rather than treating the preference parameter μ_U^i (the empowerment setpoint) as updated by *ad hoc* threshold rules, we hierarchically embed it as a latent hyperparameter subject to precision-weighted Bayesian updating.

Hierarchical Generative Model:

$$\text{Level 1: } \mu_{U,t+1}^i = \mu_{U,t}^i + \zeta_t^i, \quad \zeta_t^i \sim \mathcal{N}(0, \sigma_\mu^2) \tag{26}$$

$$\text{Level 2: } \tilde{u}_t^i \mid \mu_U^i \sim \mathcal{N}(\mu_U^i, \sigma_{\tilde{u}}^2(S_t)) \tag{27}$$

where $\tilde{u}_t^i$ is a "learning signal" derived from experienced empowerment, and $\sigma_{\tilde{u}}^2(S_t)$ is a trust-dependent observation variance (inverse precision).

4.7.2. Precision Modulation

The key mechanism is **precision modulation**: the precision (inverse variance) of the learning signal depends on current trust level S_t.

$$\Pi_{\tilde{u}}(S_t) = \sigma_{\tilde{u}}^{-2}(S_t) = \Pi_{\min} + (\Pi_{\max} - \Pi_{\min}) \cdot \sigma\big(\lambda(S_t - \theta_S)\big) \tag{28}$$

where $\sigma(\cdot)$ is the logistic function, and λ controls the sharpness of the transition.

Interpretation: When trust is high ($S_t > \theta_S$), the precision $\Pi_{\tilde{u}}$ approaches $\Pi_{\max}$, meaning observed empowerment is treated as a reliable signal for updating preferences. When trust is low, precision approaches $\Pi_{\min}$, and the same observation has minimal influence on preference learning. This implements the psychological intuition that high empowerment experiences in untrusted environments are likely to be attributed to external factors ("just happened to work out") rather than internalized as genuine capability.

Theatre workshop interpretation: A participant who spontaneously leads a scene during an early, awkward session may dismiss the experience as a fluke. The same experience in a cohesive, supportive group is more likely to be internalized as "I can do this"—updating the agent's preference setpoint upward.

4.7.3. Variational Update (Kalman Form)

Given the hierarchical model, the variational (or Kalman) update for the preference mean is:

$$\mu_U^{i,\text{post}} = \mu_U^{i,\text{prior}} + K_t^i \cdot \mathbb{E}[z_t^i] \cdot (\tilde{u}_t^i - \mu_U^{i,\text{prior}}) \tag{29}$$

where the **Kalman gain** K_t^i is determined by the precision ratio:

$$K_t^i = \frac{v_\mu \cdot \Pi_{\tilde{u}}(S_t)}{v_\mu \cdot \Pi_{\tilde{u}}(S_t) + 1} \tag{30}$$

and v_μ is the prior variance of the preference hyperparameter.

 doi:10.20944/preprints202603.0933.v1

4.7.4. Empowerment Overshoots Latent Variable

To implement the empirical observation that comfort zones "expand but do not easily contract," we introduce a latent **empowerment overshoots indicator** $z_t^i \in \{0, 1\}$:

$$\mathbb{E}[z_t^i] = \sigma\left(\lambda_z(u_t^i - \mu_U^i - \theta_{\text{gap}})\right) \tag{31}$$

When $z_t^i = 1$, the empowerment experience is treated as a "mastery" signal that updates preferences. When $z_t^i = 0$, the experience is attributed to external factors and does not update preferences.

Theatre workshop interpretation: The empowerment overshoots correspond to Grotowski's "removal of habitual defenses" [7]—a moment when the participant breaks through their usual shell and acts beyond their prior comfort zone ($u > \mu_U + \theta_{\text{gap}}$). Only such threshold-crossing experiences expand the preference setpoint. Incremental improvements within the current comfort zone do not trigger this update, reflecting the experiential distinction between routine practice and transformative breakthrough.

Unidirectional expansion: By setting $\mathbb{E}[z_t^i] = 0$ when $u_t^i < \mu_U^i$, we ensure that preferences only update upward, never downward. This implements the "ratchet effect" whereby comfort zones expand irreversibly.

Combined Update: The effective update weight is $K_t^i \cdot \mathbb{E}[z_t^i]$, combining:

- **Precision gating** (K_t^i): How much to trust the observation (trust-dependent)
- **Mastery gating** ($\mathbb{E}[z_t^i]$):
  Whether the experience qualifies as genuine growth (gap-dependent)

Table 4. Preference learning parameters.

Symbol	Description	Value
Π_{min}	Minimum learning precision (low trust)	0.01
Π_{max}	Maximum learning precision (high trust)	0.5
λ	Precision modulation sharpness	4.0
θ_S	Trust threshold for precision center	0.5
σ_μ	Hyperstate drift noise	0.005
v_μ	Prior variance of preference	0.1
λ_z	Mastery detection sharpness	5.0
θ_{gap}	Minimum gap for mastery detection	0.3

4.7.5. Relationship to Prior Work

This formulation connects to several strands of Active Inference literature:

- **Preferential Inference**: Da Costa et al. [4] define preferential inference as approximating $P(s, o|h_{\leq t})$ with $Q(s, o|h_{\leq t})$, where the preference model depends on history. Our extension treats preference *parameters* (not just the conditional distribution) as subject to inference.
- **Precision as confidence**: The interpretation of precision as "confidence" or "reliability" is standard in Active Inference [34]. Here, we apply this principle to preference learning rather than just observation.
- **Empirical priors**: Friston's formulation of empirical priors [3] allows priors to depend on random variables learned from experience. Our preference hyperparameter μ_U functions as such an empirical prior.

4.8. EFE Computation and Action Selection

4.8.1. State Variable Prediction

Prediction of each state variable τ steps ahead. Note that **the agent's predictive model is a simplification of the true environmental dynamics**. In particular, the trust–empowerment coupling

 doi:10.20944/preprints202603.0933.v1

term $\kappa_{S\to u} \cdot \sigma(S_t) \cdot \tilde{e}_t^{(i)}$ that appears in the environmental state transitions (Eq. 12) is omitted from the agent's internal predictive model. This design choice reflects three considerations:

1. **Partial observability of trust**: Since S_t is only partially observable, the agent cannot condition predictions on its true value. Substituting the belief mean m_S^i would add uncertainty and bookkeeping to multi-step predictions.
2. **Complexity of social interaction**: Predicting co-evolving others' behavior and collective trust is cognitively demanding. Omitting this interaction term is consistent with bounded rationality: agents deploy a simpler internal model that respects realistic limits on prospective social prediction [35].
3. **Conservative prediction**: Without the trust-mediated bonus, predicted empowerment gains are conservative; any such gains realized in the environment appear as positive surprise.

This *model mismatch* is consistent with bounded rationality [35] and does not prevent effective action selection.

(a) Collective Trust S (Partially Observable $\to$ Expressed as Belief)

$$Q(S_{t+\tau} \mid a, h^i_{\le t}) = \mathcal{N}(m^a_{t+\tau}, v^a_{t+\tau}) \tag{32}$$

where mean and variance are computed recursively:

$$m^a_{t+\tau} = a_S m^a_{t+\tau-1} + b_S + c_S^{\text{Hill}} \mathcal{H}_n(\hat{e}; K) \tag{33}$$

$$v^a_{t+\tau} = a_S^2 v^a_{t+\tau-1} + \sigma_S^2 \tag{34}$$

where $\hat{e} = \mathbf{1}[a = 2]$ is the expression indicator for action a.

(b) Empowerment u (Fully Observable $\to$ Point Estimate)

The agent's predictive model for empowerment omits the trust–empowerment coupling:

$$\begin{aligned} \hat{u}^{i,a}_{t+\tau} = {} & \alpha_U \hat{u}^{i,a}_{t+\tau-1} + \eta_{\text{self}} \cdot \mathbf{1}[a = 2] \\ & + \eta_{\text{other}} \cdot \hat{\tilde{e}}^{-i} \cdot \left(1 + \gamma_{\text{coop}} \cdot \sigma(\hat{u}^{i,a}_{t+\tau-1})\right) \end{aligned} \tag{35}$$

where $\mathbf{1}[a = 2]$ indicates that only Express actions contribute self-expression gains.

(c) Stamina H (Fully Observable $\to$ Point Estimate)

$$\hat{H}^{i,a}_{t+\tau} = \begin{cases} \min\left(\hat{H}^{i,a}_{t+\tau-1} + H_{\text{rec}}, H_{\max}\right) & \text{if } a = 0 \text{ (Rest)} \\ \hat{H}^{i,a}_{t+\tau-1} - c_{\text{chat}} & \text{if } a = 1 \text{ (Chat \& Exercise)} \\ \max\left(\hat{H}^{i,a}_{t+\tau-1} - c_{\text{exp}} \cdot (1 - c_H^{\text{Hill}} \cdot \mathcal{H}_{n_H}(\hat{\tilde{e}}^{-i}; K_H)), 0\right) & \text{if } a = 2 \text{ (Express)} \end{cases} \tag{36}$$

4.8.2. Risk Computation (All Four State Variables)

Risk is the KL divergence between predicted states and preferred states. Assuming independence of each state variable:

$$\begin{aligned} \text{Risk}_{t+\tau}(a) = {} & \text{Risk}^S_{t+\tau}(a) + \text{Risk}^u_{t+\tau}(a) \\ & + \text{Risk}^H_{t+\tau}(a) + \text{Risk}^W_{t+\tau}(a) \end{aligned} \tag{37}$$

(a) Risk for Trust (KL divergence between Gaussian distributions) We quantify pragmatic deviation from preferences by the Kullback–Leibler divergence between the agent's predicted belief over $S_{t+\tau}$ and its preferred distribution. Let

$$q(S_{t+\tau}) = \mathcal{N}(m^a_{t+\tau}, v^a_{t+\tau}), \qquad p(S_{t+\tau}) = \mathcal{N}(\mu^i_S, \sigma^2_{\text{pref},S}). \tag{38}$$

 doi:10.20944/preprints202603.0933.v1

Then $\text{Risk}^S_{t+\tau}(a) = D_{\text{KL}}[q \,\|\, p]$ admits a closed form. Starting from $D_{\text{KL}}[q \,\|\, p] = \mathbb{E}_q[\log q - \log p]$ and using the standard Gaussian identities for quadratic expectations,

$$\mathbb{E}_q[(S - \mu^i_S)^2] = (m^a_{t+\tau} - \mu^i_S)^2 + v^a_{t+\tau}, \tag{39}$$

we obtain the expression below.

$$\begin{aligned} \text{Risk}^S_{t+\tau}(a) &= D_{\text{KL}}\Big[\mathcal{N}(m^a_{t+\tau}, v^a_{t+\tau}) \,\|\, \mathcal{N}(\mu^i_S, \sigma^2_{\text{pref},S})\Big] \\ &= \frac{1}{2}\left[\frac{v^a_{t+\tau}}{\sigma^2_{\text{pref},S}} + \frac{(m^a_{t+\tau} - \mu^i_S)^2}{\sigma^2_{\text{pref},S}} - 1 - \log\frac{v^a_{t+\tau}}{\sigma^2_{\text{pref},S}}\right] \end{aligned} \tag{40}$$

Implementation note: For computational efficiency, the implementation uses a quadratic approximation that omits the constant and logarithmic terms:

$$\text{Risk}^S_{t+\tau}(a) \approx w_S\Big[(m^a_{t+\tau} - \mu^i_S)^2 + v^a_{t+\tau}\Big] \tag{41}$$

where w_S is a weighting coefficient. This approximation is motivated by retaining only the terms that vary with action-dependent predictions: the mean-deviation penalty $(m^a_{t+\tau} - \mu^i_S)^2$ and the uncertainty penalty $v^a_{t+\tau}$. The remaining terms (-1 and the logarithmic ratio) are either constant under fixed preference variance or can introduce numerical instability when $v^a_{t+\tau}$ becomes very small.

(b) Risk for Empowerment

Empowerment u is treated as (effectively) fully observable to the agent, so the pragmatic term does not require a belief distribution with substantial uncertainty. Formally, one can recover the squared-deviation form as a small-variance limit of a KL divergence. For example, represent the predicted state by a narrow Gaussian $q(u_{t+\tau}) = \mathcal{N}(\hat{u}^{i,a}_{t+\tau}, \epsilon)$ with $\epsilon \to 0$, and the preference by $p(u_{t+\tau}) = \mathcal{N}(\mu^i_U, \sigma^2_{\text{pref},U})$. Then

$$D_{\text{KL}}[q \,\|\, p] = \frac{1}{2}\left[\frac{(\hat{u}^{i,a}_{t+\tau} - \mu^i_U)^2 + \epsilon}{\sigma^2_{\text{pref},U}} - 1 - \log\frac{\epsilon}{\sigma^2_{\text{pref},U}}\right], \tag{42}$$

so, up to constants and a rescaling, Risk reduces to a squared deviation from the preferred empowerment level. Accordingly, we use:

$$\text{Risk}^u_{t+\tau}(a) = \frac{w_U}{2}\left(\hat{u}^{i,a}_{t+\tau} - \mu^i_U\right)^2 \tag{43}$$

Implementation note: The implementation includes a predictive variance term to account for future uncertainty in u:

$$\text{Risk}^u_{t+\tau}(a) \approx w_U\left[\left(\hat{u}^{i,a}_{t+\tau} - \mu^i_U\right)^2 + v^{\text{prior}}_{u,t+\tau}\right] \tag{44}$$

where $v^{\text{prior}}_{u,t+\tau}$ accumulates process noise over the prediction horizon.

extbf(c) Risk for Stamina

$$\text{Risk}^H_{t+\tau}(a) = \frac{w_H}{2}\left(\hat{H}^{i,a}_{t+\tau} - \mu^i_H\right)^2 \tag{45}$$

This form corresponds to the "fully observed" or "zero-variance" limit: if one were to represent stamina with a narrow Gaussian belief $\mathcal{N}(\hat{H}^{i,a}_{t+\tau}, \epsilon)$ and take $\epsilon \to 0$, the KL-based risk reduces (up to constants) to a squared deviation from the preferred level μ^i_H.

extitImplementation note: Similarly to empowerment, the implementation can include a predictive variance term for H to penalize uncertainty accumulation over the horizon:

$$\text{Risk}^H_{t+\tau}(a) \approx w_H\left[\left(\hat{H}^{i,a}_{t+\tau} - \mu^i_H\right)^2 + v^{\text{prior}}_{H,t+\tau}\right], \tag{46}$$

 doi:10.20944/preprints202603.0933.v1

where $v^{\text{prior}}_{H,t+\tau}$ summarizes accumulated process noise in the stamina prediction.

(d) Risk for Interpersonal Trust (aggregated over neighborhood)

The Risk for interpersonal trust is computed using the **aggregated interpersonal trust** $\bar{W}^{(i)}$ and its uncertainty $\bar{v}^{(i)}_W$:

$$\bar{W}^{(i)}_t = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} m_{W_{ij},t} \tag{47}$$

$$\bar{v}^{(i)}_{W,t} = \frac{1}{|\mathcal{N}(i)|^2} \sum_{j \in \mathcal{N}(i)} v_{W_{ij},t} \tag{48}$$

where $m_{W_{ij}}$ and $v_{W_{ij}}$ are the mean and variance of agent i's belief about W_{ij}.

The variance expression follows from a standard propagation-of-uncertainty argument. If the dyadic beliefs are treated as conditionally independent Gaussians $W_{ij} \sim \mathcal{N}(m_{W_{ij},t}, v_{W_{ij},t})$ given $h^i_{\leq t}$, then their average

$$\bar{W}^{(i)}_t = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} W_{ij,t} \tag{49}$$

is also Gaussian with

$$\begin{aligned} \text{Var}(\bar{W}^{(i)}_t) &= \frac{1}{|\mathcal{N}(i)|^2} \sum_{j \in \mathcal{N}(i)} \text{Var}(W_{ij,t}) \\ &= \frac{1}{|\mathcal{N}(i)|^2} \sum_{j \in \mathcal{N}(i)} v_{W_{ij},t}. \end{aligned} \tag{50}$$

which yields $\bar{v}^{(i)}_{W,t}$.

The Risk penalizes deviation from a neutral interpersonal trust preference μ_W:

$$\text{Risk}^W_{t+\tau}(a) = w_W \left[\left(\bar{W}^{(i)}_{t+\tau} - \mu_W \right)^2 + \bar{v}^{(i)}_{W,t+\tau} \right] \tag{51}$$

This quadratic form can be viewed as the same approximation used for collective trust (Eq. 41), applied to the aggregated latent variable $\bar{W}^{(i)}$. Intuitively, it encourages neighborhood-averaged trust to stay near a "neutral" target while discouraging policies that leave large uncertainty in dyadic relations.

Parameters:

- $w_W = 0.5$: Interpersonal trust weight in EFE
- $\mu_W = 0.5$: Preferred interpersonal trust (neutral)

4.8.3. Information Gain Computation (Computed for latent S and W)

In our formulation, the information-gain (epistemic) term is computed for **partially observable latent variables**: the collective trust S and the interpersonal trust W_{ij}. Because (u, H) are directly observed at each time step in the model, they yield *zero observation-based* information gain under the assumed observation model. Importantly, this does *not* mean that epistemic value is irrelevant for predicting future internal dynamics: since the transitions of u and H depend (directly or indirectly) on S and W, reducing uncertainty about these latent variables also reduces uncertainty about *future* u and H trajectories through the predictive model.

Under the linear Gaussian model, information gain for S can be computed exactly:

$$I^S_{t+\tau}(a) = I(S_{t+\tau}; o^i_{t+\tau} \mid a) = \frac{1}{2} \log \left(1 + \frac{H^2 v^a_{t+\tau-1}}{R(a)} \right) \tag{52}$$

where H is the observation matrix (linearized gradient of the observation function).

 doi:10.20944/preprints202603.0933.v1

Implementation note: The implementation assumes a direct observation model with $H = 1$, yielding:

$$I^S_{t+\tau}(a) \approx \frac{1}{2}\log\left(1 + \frac{v^a_{t+\tau-1}}{R(a)}\right) \tag{53}$$

This simplification is appropriate when the observation function is approximately linear near the current belief mean.

Important Design: Since $R(a = 2) < R(a = 0)$ (expressing reduces observation noise), expression action increases information gain for S.

Information Gain for Interpersonal Trust W (Chat & Exercise only)

A key design feature of the network model is that **information gain about interpersonal trust W_{ij} is obtained exclusively through Chat & Exercise actions**. This separation implements the distinction between:

- **Curiosity-driven social exploration** (Chat & Exercise): Low-commitment interaction that reveals information about interpersonal relationships.
- **Expression-driven coordination** (Express): High-commitment action focused on building collective trust and empowerment.

The information gain for interpersonal trust is:

$$I^W_{t+\tau}(a) = \begin{cases} \frac{1}{2}\log\left(1 + \frac{\bar{v}^{(i)}_{W,t+\tau}\,|\mathcal{N}(i)|}{\sigma^2_W}\right) & a = 1 \\ 0 & \text{otherwise} \end{cases} \tag{54}$$

where $\bar{v}^{(i)}_{W,t+\tau}$ is the average variance of beliefs about neighbors' trust. Note that the term is scaled by $|\mathcal{N}(i)|$, reflecting that Chat & Exercise provides information about *all* neighbors simultaneously. While this implies that agents with higher degrees (more connections) can potentially gain more total information, this bias is consistent with the social reality that well-connected individuals have more to learn from social interaction. In our simulations using Erdős-Rényi graphs, the degree distribution is relatively homogeneous, minimizing the impact of this potential bias.

Interpretation: Chat & Exercise provides observations that reduce uncertainty about neighbors' trustworthiness ($\text{IG}_W > 0$), while Rest and Express do not provide such information ($\text{IG}_W = 0$). However, *both Chat & Exercise and Express update the true value of W_{ij}* through action synchrony (Eq. 15): Express–Express synchrony strongly builds interpersonal trust ($\phi = +1.0$), while Chat & Exercise synchrony moderately builds trust ($\phi = +0.5$).

This creates a three-way tradeoff in action selection:

- **Rest**: Stamina recovery, minimal information gain, weak synchrony effect on W_{ij}.
- **Chat & Exercise**: Moderate stamina cost, $\text{IG}_W > 0$ (curiosity-driven exploration), moderate synchrony effect on W_{ij}.
- **Express**: High stamina cost, $\text{IG}_S > 0$ (collective trust observation), empowerment gain, *strong* synchrony effect on W_{ij} but $\text{IG}_W = 0$.

Theatre workshop interpretation: Joint improvisation (Express) builds solidarity through shared creative activity—interpersonal trust W_{ij} genuinely increases—but participants are absorbed in the performance itself, not attending to how others perceive them. Only through conversation and group exercises (Chat & Exercise) can an agent recognize, "Ah, that person trusts me." This asymmetry between *building* trust and *knowing* about trust captures a realistic feature of experiential learning.

4.8.4. Differing Roles of Fully and Partially Observable Variables

This structure implies:

- For S: Both information gain (via Express) and risk contribute.

 doi:10.20944/preprints202603.0933.v1

Table 5. Observability and EFE contributions of state variables.

Variable	Observability	Contribution to Risk	Information Gain
S (Collective Trust)	Partial	Belief vs. preference	Yes (Express)
W_{ij} (Interpersonal Trust)	Partial	Aggregated belief vs. preference	Yes (Chat & Ex. only)
u (Empowerment)	Full	Prediction vs. preference	No
H (Stamina)	Full	Prediction vs. preference	No

- For W_{ij}: Information gain (via Chat & Exercise) and risk contribute; Chat & Exercise enables curiosity-driven social exploration.
- For u, H: Risk contributes directly; epistemic drive enters *indirectly* via reduced uncertainty in S and W which improves multi-step prediction of (u, H).

4.8.5. N-Step Expected Free Energy

The N-step EFE is defined as:

$$G^{(N)}(a) = \sum_{\tau=1}^{N} \Big[\underbrace{\text{Risk}^{S}_{t+\tau} + \text{Risk}^{u}_{t+\tau} + \text{Risk}^{H}_{t+\tau} + \text{Risk}^{W}_{t+\tau}}_{\text{Risk (4 vars)}} \Big] - \sum_{\tau=1}^{N} \underbrace{\left(I^{S}_{t+\tau} + I^{W}_{t+\tau}\right)}_{\text{Info Gain}} \tag{55}$$

Implementation note: When computing the N-step EFE, the variance at step τ depends on expected observations at earlier steps. The exact update requires a Kalman update at each step; for computational efficiency, the implementation uses an exponential approximation:

$$v^{\text{post}}_{t+\tau} \approx v^{\text{prior}}_{t+\tau} \cdot \exp\left(-2 \cdot I_{t+\tau}(a)\right) \tag{56}$$

This approximation exploits the relationship between information gain and variance reduction in Gaussian models.

4.8.6. Action Selection Rule

Action selection is based purely on EFE with softmax policy over the three-valued action space:

$$P(a^{i}_{t} = a) = \frac{\exp\left(-\beta_{\text{action}} \cdot G^{(N)}(a)\right)}{\sum_{a' \in \{0,1,2\}} \exp\left(-\beta_{\text{action}} \cdot G^{(N)}(a')\right)} \tag{57}$$

This is **softmax action selection**, where β_{action} is the inverse temperature parameter controlling the exploration-exploitation tradeoff. Higher β_{action} leads to more deterministic selection of the action with lowest EFE.

4.9. Propagation of Nonlinear Effects to Risk

Although the nonlinearities in this model are implemented in the state transitions (Hill-type collective effects and trust-gated learning), their behavioral consequence is expressed through the Risk component of EFE because Risk scores the mismatch between predicted trajectories and preferred setpoints. In particular, the empowerment Risk (Eq. for Risk^{u}) depends on the predicted future empowerment $\hat{u}^{i,a}_{t+\tau}$ under each candidate action. When collective trust S is high in the environment, the true dynamics provide a stronger vicarious gain channel through the trust–empowerment coupling term in Eq. 12. As a result, policies that include Express (directly or indirectly by inducing others' expression) tend to yield higher realized empowerment than in low-trust regimes. Even though the agent's internal predictive model is intentionally simplified (Section 4.8), this regime dependence still

 doi:10.20944/preprints202603.0933.v1

feeds back into planning via the accumulated prediction error and the subsequent updates of belief and preference parameters.

Comfort-zone expansion then changes the geometry of the same Risk term. Preference learning updates the empowerment setpoint μ^i_U upward when a high empowerment experience is encountered under sufficient inferred trust (Section 4.7). Because $\text{Risk}^u_{t+\tau}(a)$ penalizes deviations of $\hat{u}^{i,a}_{t+\tau}$ from μ^i_U, an increase in μ^i_U immediately reweights the EFE landscape: trajectories that would previously be "too high" in u become less risky, making sustained high-expression/high-empowerment regimes more self-consistent under EFE minimization. Conversely, in low-trust phases the precision-gating mechanism suppresses updates of μ^i_U, so the agent retains a lower setpoint and high-u trajectories remain costly.

The interaction of these two mechanisms strengthens a key signature of chaotic itinerancy: prolonged residence near multiple quasi-stable (metastable) regimes and irregular transitions among them. High-trust episodes make it easier for expression to generate large empowerment excursions; if such excursions are consolidated into a higher μ^i_U, future expressive policies become less penalized by Risk and therefore more likely to be selected, stabilizing a high-activity mode. If the system remains in a low-trust region long enough, the same consolidation does not occur, and the low-activity mode remains comparatively stable. Thus, nonlinear transition structure and trust-gated preference plasticity jointly shape the Risk term so that the group can dwell near quasi-stable regimes for extended periods and then transition to other regimes as conditions shift, consistent with the defining "residence-and-switching" pattern of CI.

5. Numerical Results

This section presents the primary numerical evidence for our model. We prioritize the **chaotic itinerancy (CI) test** as the central dynamical signature: trajectories that dwell near multiple quasi-stable "attractor ruins" and switch irregularly among them. CI provides the most complete characterization of the model's complex dynamics, capturing the structured variability observed in real workshop settings.

As supplementary verification, we also report:

- **Multiple-equilibria analysis**: Demonstrating that multiple metastable regimes coexist, though not necessarily as strongly separated bistable attractors.
- **Intervention response check**: Confirming that perturbations produce distinct trajectories, a natural consequence of itinerant dynamics among multiple regimes.

The sensitivity analysis reveals the following key findings:

- **Hill coefficient** n: Bistability requires $n \geq 4$, and CI is robust across all bistable values (70–85%). The default $n = 4$ achieves 85% CI pass rate while remaining within the biologically plausible range for cooperativity indices.
- **Half-saturation constant** K: Both bistability and CI depend sensitively on K. The bistable window spans $K \in [0.27, 0.70]$, with CI pass rates of 70–90% in the range $K \in [0.20, 0.60]$. The default $K = 0.40$ achieves robust CI (80%) at the center of the bistable region.
- **Precision modulation** λ: CI is remarkably robust across the entire tested range ($\lambda \in [2, 16]$), with pass rates of 80–90%. This indicates that the precision-gated learning mechanism is not sensitive to the specific sharpness of the trust threshold.
- **Coupling strength** $\kappa_{S\to u}$: This parameter is the primary driver of both bistability and CI. CI occurrence increases monotonically with coupling strength, from 70% at $\kappa = 0.2$ to 100% at $\kappa = 0.8$. The default $\kappa = 0.6$ balances robust CI with moderate multistability.

In summary, the model's chaotic itinerancy is robust across a wide range of parameter values, with the trust–empowerment coupling strength acting as the primary control parameter.

5.1. Primary Test: Chaotic Itinerancy—Metastable Switching Among Attractor Ruins

The primary dynamical signature we target is *chaotic itinerancy* (CI): trajectories that dwell near multiple quasi-stable "attractor ruins" and switch irregularly among them [10–12,36]. Unlike simple bistability, CI emphasizes *structured variability*: the system neither converges to a single fixed point nor oscillates periodically. Instead, trajectories exhibit irregular residence times near multiple quasi-stable regimes, with transitions sensitive to both noise and intervention. This framework provides an interpretable vocabulary for small-group processes where collective modes (e.g., inhibited, playful, exploratory, confrontational) transiently stabilize and then reorganize.

 doi:10.20944/preprints202603.0933.v1

Default parameter verification.

At the default parameters ($N = 6$, $n = 4$, $\sigma_S = 0.03$, seed=1, $T = 400$), the model passes all three primary CI indicators grounded in the theoretical literature:

1. **Multiple attractor ruins** [37]: 8 distinct metastable clusters detected via HDBSCAN.
2. **Heavy-tailed residence times** [38]: Skewness = 14.04 $>$ 1.5, indicating structured dwelling near quasi-stable regimes.
3. **Local splitting exponent signature** [39–41]: Mean $\approx$ 0 (0.025), substantial variance ($\sigma = 0.128 >$ 0.05), and frequent sign changes (27.3% $>$ 20%), consistent with alternating local stability and instability.

These three indicators directly operationalize the defining characteristics of chaotic itinerancy established in the theoretical physics literature.

To interpret the FTLE panel used in our diagnostic plots, note that a finite-time Lyapunov exponent (FTLE) is a local measure of **how quickly nearby trajectories separate** over a finite time window. In practice we estimate an FTLE proxy from the simulated trajectory embedded in a low-dimensional feature space (the same representation used for clustering into attractor ruins), and we compute a finite-window **local divergence rate**. Positive values indicate local expansion (instability; sensitivity to perturbations), whereas negative values indicate local contraction (stability; dwelling near an attractor ruin). Chaotic itinerancy is characterized by near-marginal stability: the FTLE fluctuates around zero with substantial variance and frequent sign changes, reflecting alternation between quasi-stable residence and transiently unstable transition epochs.

Figure 1 presents the comprehensive CI diagnostic panel for the default parameter configuration.

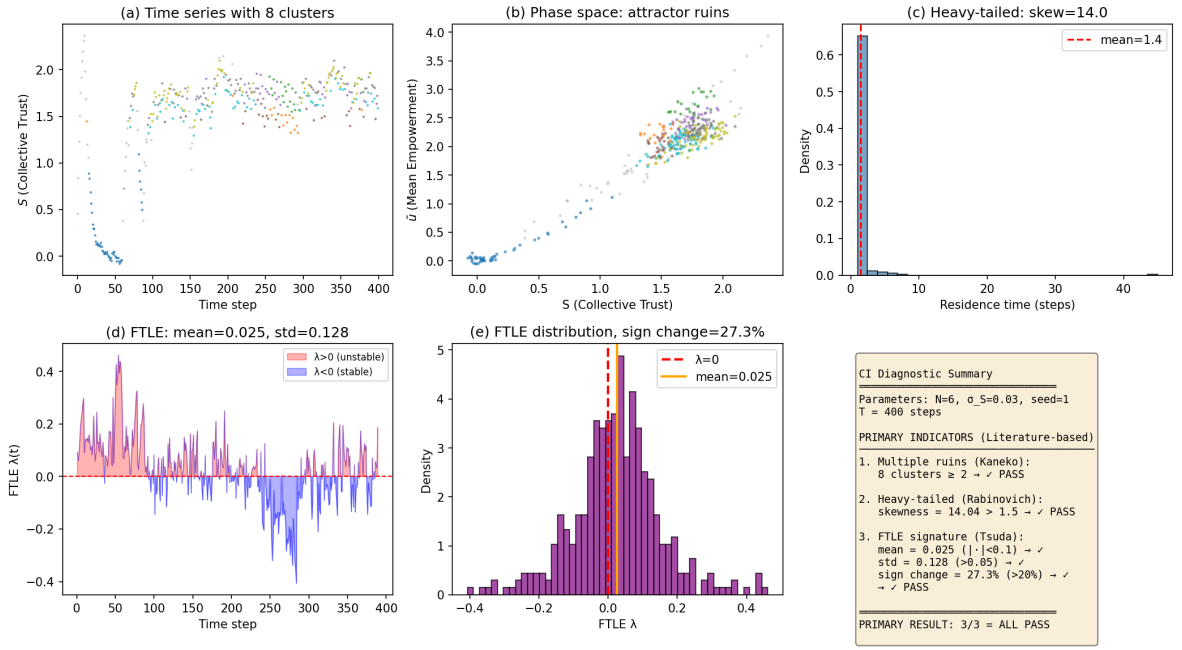

Figure 1. Chaotic itinerancy diagnostic panel under default parameters ($N = 6$, $n = 4$, $\sigma_S = 0.03$, seed=1). The six-panel display shows: (top-left) time series of collective trust S and expression rate $\bar{e}$; (top-middle) phase space trajectory with color indicating time progression; (top-right) residence time distribution showing heavy-tailed structure; (bottom-left) FTLE time series with sign changes indicating alternating stability; (bottom-middle) cluster membership over time; (bottom-right) action distribution across agents.

Figure 2 shows the long-term chaotic itinerancy dynamics, demonstrating the irregular switching between high and low expression states.

 doi:10.20944/preprints202603.0933.v1

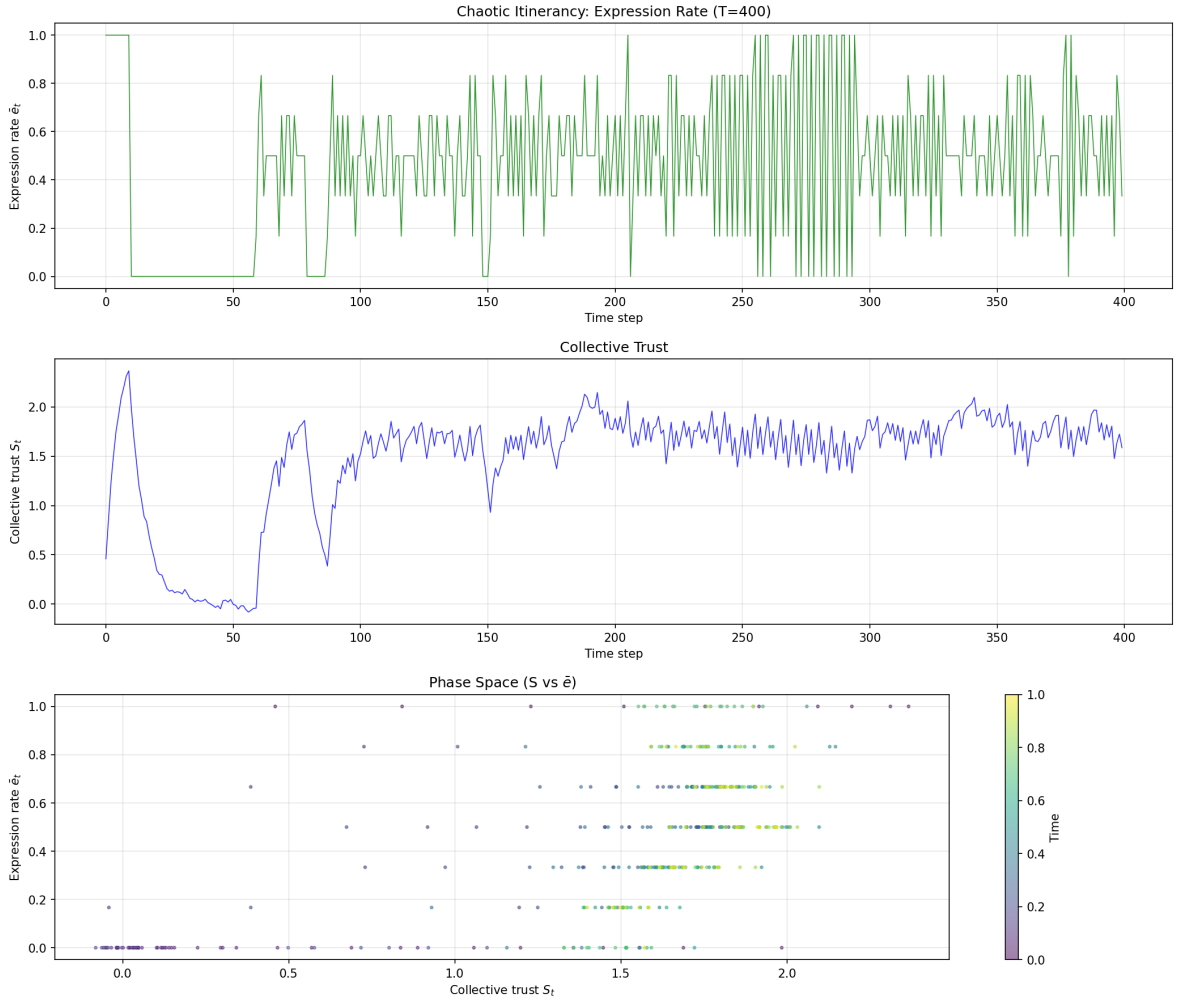

Figure 2. Chaotic itinerancy dynamics over 400 time steps. (Top) Expression rate $\bar{e}_t$ exhibiting irregular oscillations between high and low states. (Middle) Collective trust S_t showing correlated fluctuations. (Bottom) Phase space plot (S vs $\bar{e}$) with color indicating time, revealing the trajectory's visits to multiple quasi-stable regions.

 doi:10.20944/preprints202603.0933.v1

Parameter scan (overview).

To assess robustness across parameter space, we conducted a parameter scan varying three control parameters. We ran $T = 400$ time steps with $N = 20$ agents and performed 3 trials per configuration. The scanned grid comprised $\beta_{\text{action}} \in \{1, 2, 3, 5, 8\}$, $n \in \{4, 6, 8, 10\}$, and trust-state noise $\sigma_S \in \{0.02, 0.05, 0.08, 0.12\}$ (80 configurations total).

Each trial was evaluated using five indicators: the three primary CI indicators above, plus two supplementary operational indicators (transition-state fraction and action variability) that help identify promising parameter regions. The full indicator definitions and thresholds are reported in Supplementary Material (Sec. S-IV).

Results.

Across the 80 scanned configurations, 63 (78.8%) were classified as CI regions. Marginally, CI was most prevalent at intermediate trust noise ($\sigma_S = 0.05$: 95% CI; other values: 70–80%), and increased with higher action precision ($\beta_{\text{action}} = 8$: 87.5% CI). Across Hill exponents, CI fractions were high for $n = 4$ and $n = 10$ (both 85%), and lower at $n = 6$ (70%), suggesting that the degree of nonlinearity shapes whether trajectories become trapped or continue to switch among metastable regimes.

Table 6 reports the marginal CI fractions and mean indicator scores, averaged over the other scan dimensions.

Table 6. Chaotic itinerancy (CI) marginal results in the parameter scan (80 configurations; 3 trials each). "CI fraction" is the fraction of configurations classified as CI within each marginal slice; "mean indicators" is the average of the five-indicator score.

Parameter	Value	CI fraction	Mean indicators
β_{action}	1.0	81.2%	2.69
β_{action}	2.0	68.8%	2.62
β_{action}	3.0	81.2%	2.77
β_{action}	5.0	75.0%	2.77
β_{action}	8.0	87.5%	2.79
n	4	85.0%	2.68
n	6	70.0%	2.65
n	8	75.0%	2.75
n	10	85.0%	2.83
σ_S	0.02	70.0%	2.63
σ_S	0.05	95.0%	2.93
σ_S	0.08	80.0%	2.72
σ_S	0.12	70.0%	2.63

These findings provide an additional robustness check beyond two-attractor multistability. They suggest that, for a broad parameter set, the same generative mechanisms that yield multistability and learning-induced hysteresis can also generate structured metastable switching compatible with CI-style dynamics, rather than requiring fine-tuned deterministic chaos.

Figure 3 shows the distribution of actions across the agent population, illustrating the heterogeneity in behavioral choices.

5.2. Supplementary Check 1: Multiple-Equilibria Analysis

As a supplementary verification, we examined whether the model exhibits multiple coexisting metastable regimes. Note that chaotic itinerancy does not require strong bistability; rather, it involves quasi-stable "attractor ruins" that trajectories visit transiently. Nevertheless, demonstrating multiple equilibria provides additional evidence for the model's capacity to support structured collective dynamics.

Importantly, this test is conceptually different from the CI analysis. The CI test focuses on a **single simulation run** and asks whether, within one time-series path, the trajectory exhibits repeated residence near quasi-stable regimes and irregular transitions among them. In contrast, the multiple-

 doi:10.20944/preprints202603.0933.v1

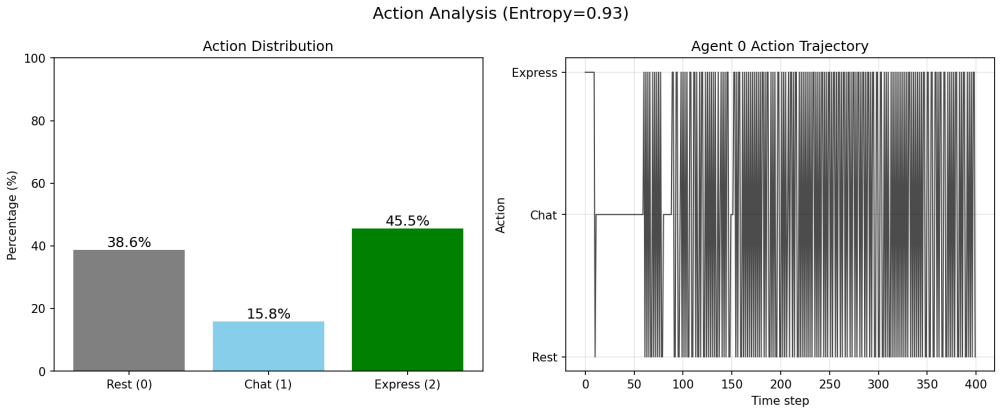

Figure 3. Action distribution across agents during CI dynamics. The distribution shows the relative frequency of Rest (0), Chat & Exercise (1), and Express (2) actions, revealing the collective behavioral patterns that emerge from individual active inference processes.

equilibria test deliberately runs **many simulations with different initial conditions** under identical parameters, and then compares their **time-averaged states** (e.g., S^* and $\bar{e}^*$) to assess whether the dynamics support multiple distinct outcomes. In other words, CI emphasizes within-trajectory switching among metastable modes, whereas the multiple-equilibria test emphasizes across-run separation of long-run averages as evidence for more than one equilibrium.

We simulated $N = 6$ agents for $T = 400$ time steps with parameters listed in Table S1 in the Supplementary Material. Starting from identical model parameters but different initial conditions, the system converges to distinct steady states.

Definition (Separation Degree): The *separation degree* quantifies the magnitude of multistability in the two-attractor (bistable) case by measuring the difference in equilibrium trust levels between trajectories converging to different attractors:

$$\Delta_{\text{sep}} := S^*_{\text{high}} - S^*_{\text{low}} \tag{58}$$

where S^*_{high} and S^*_{low} denote the time-averaged trust levels over the final 50 time steps for high and low attractors, respectively. A separation degree exceeding 1.0 indicates that the system exhibits meaningful multistability with well-separated attractors.

Basin of Attraction Analysis: To systematically characterize multistability, we employed a grid-based basin of attraction mapping. Initial conditions were sampled from a 4×4 grid spanning $(S_0, u_0) \in [0.1, 0.9] \times [0.1, 0.9]$, yielding 16 independent trajectories. Each trajectory was evolved for $T = 400$ steps to ensure convergence to steady state, and the final expression rate $\bar{e}^*$ was recorded as the time-average over the final 30 steps. Trajectories were classified as "high-expression" if $\bar{e}^* > 0.2$ and "low-expression" otherwise.

Table 7 summarizes the basin analysis for the default parameter configuration ($n = 4$).

Table 7. Basin of attraction analysis demonstrating multistability ($n = 4$, default parameters)

Attractor	Count (of 16)	Mean S^*	Mean $\bar{e}^*$
High-expression	12	1.63	0.51
Low-expression	4	−0.03	0.00
Separation degree: $\Delta_{\text{sep}} = 1.65$			

The separation degree of 1.65 demonstrates clear multistability with well-separated attractors. The Gaussian latent variable formulation allows trust to take values significantly above 1 (high-expression attractor at $S^* \approx 1.6$), reflecting the theoretical consistency of unbounded state evolution.

Key observations:

 doi:10.20944/preprints202603.0933.v1

- **Two distinct attractors**: The system exhibits two distinct states in average, with the high-expression attractor characterized by $\bar{e}^* \approx 0.51$ and the low-expression attractor by $\bar{e}^* \approx 0.00$.
- **Asymmetric basins**: The high-expression attractor captures 12/16 (75%) of initial conditions, indicating that the system tends toward collective participation under most starting conditions.
- **Robust state separation**: Trust levels differ by $\Delta S \approx 1.7$ between attractors, indicating robust separation that is unlikely to be bridged by noise fluctuations alone.

5.3. Supplementary Check 2: Intervention Response

A natural consequence of chaotic itinerancy is that perturbations can redirect the system's trajectory among metastable regimes. We checked this *intervention responsiveness* by comparing trajectories under different perturbation histories.

Note that this differs from classical "path dependence" in bistable systems, where trajectories deterministically lock into one of two attractors. In a CI regime, trajectories may continue to wander among multiple regimes; the key observation is that different intervention histories produce distinguishable trajectory ensembles.

To demonstrate intervention responsiveness, we conducted an *intervention pulse experiment* inspired by the discussion of basin stability in [42]. Two scenarios were compared starting from **identical initial conditions** ($S_0 = 0.0$, $\bar{e}_0 = 0.5$), averaged over 5 random seeds (seeds 1–5):

- **High pulse**: Trust is externally set to $S = 0.9$ at $t = 50$
- **Low pulse**: Trust is externally set to $S = 0.1$ at $t = 50$

Table 8. Intervention response check: Divergence under different pulse interventions (5-seed average)

Condition	Mean $\bar{e}$ (post-intervention)	Std
High pulse ($S = 0.9$)	0.39	0.20
Low pulse ($S = 0.1$)	0.29	0.24
Divergence magnitude: $\lvert\Delta\bar{e}_{\text{avg}}\rvert = 0.10$		

This demonstrates intervention responsiveness: starting from *identical initial conditions* ($S_0 = 0.0$, $\bar{e}_0 = 0.5$), different intervention pulses lead to distinguishable trajectory ensembles when averaged over multiple random seeds. The divergence magnitude of approximately 0.10 indicates that the high pulse biases the system toward higher expression levels, consistent with the expected effect of trust-building interventions.

The intervention triggers a cascade of effects:

1. **Trust exceeds precision threshold**: During the high pulse, $S > \theta_S$ activates high-precision preference learning.
2. **Empowerment exceeds preference**: The enhanced environment produces empowerment gains that exceed current preferences.
3. **Comfort zone expansion**: Precision-gated preference learning updates μ_U upward.
4. **Irreversible preference shift**: The elevated preference persists after the intervention ends.
5. **Convergence to high attractor**: The system converges to the high-expression steady state.

This supports a learning-induced mechanism: the intervention leaves a lasting trace in agents' preference parameters (via μ_U updates), not only in state trajectories. Importantly, this differs from classical path dependence in strongly bistable systems: in a CI regime, the system's future remains open, with interventions shaping the sequence of visited modes rather than determining a single final state—consistent with the theatrical intuition that workshops are exploratory processes where outcomes emerge from ongoing interaction.

Figure 4 demonstrates the effect of facilitator intervention on system trajectories.

 doi:10.20944/preprints202603.0933.v1

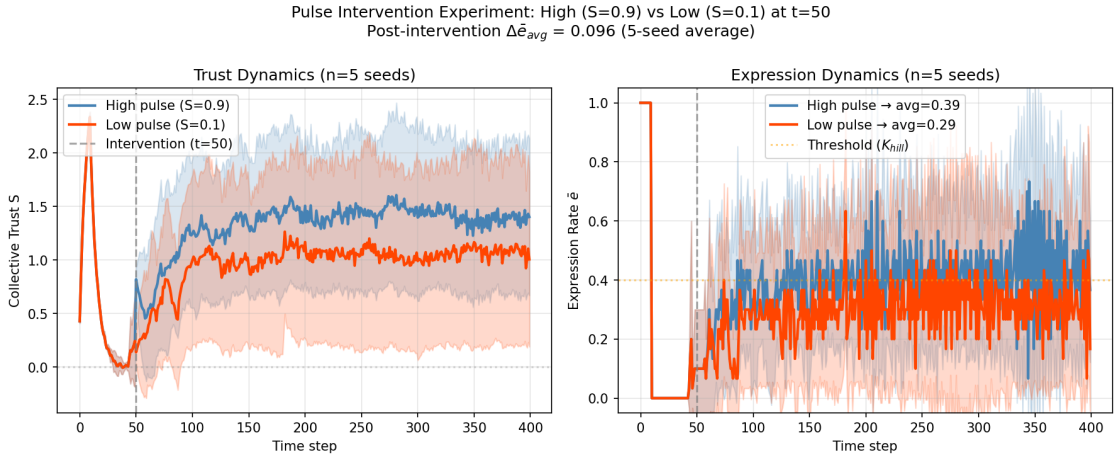

Figure 4. High vs. Low pulse intervention experiment demonstrating intervention responsiveness (5-seed average with standard deviation bands). Both trajectory ensembles start from identical initial conditions ($S_0 = 0.0, \bar{e}_0 = 0.5$) and receive a single pulse at $t = 50$. The **high pulse** ($S \to 0.9$, blue) biases the system toward higher expression levels, while the **low pulse** ($S \to 0.1$, red) produces lower average expression. The divergence ($\Delta\bar{e}_{\text{avg}} \approx 0.10$) demonstrates that facilitator interventions can redirect collective dynamics, consistent with therapeutic and workshop practice.

5.4. Phase Transition Characteristics

We next characterize how the nonlinear collective effect shapes the system-level attractor landscape. Recall that the trust dynamics include a Hill-type cooperative term $\mathcal{H}_n(\bar{e}; K)$, where K is the *half-saturation constant* (critical threshold). Operationally, K specifies the collective expression level $\bar{e}$ at which the cooperative amplification of trust becomes effective. Therefore, K acts as a control parameter: small K makes the amplification easy to trigger (biasing the system toward high activity), large K makes it difficult (biasing toward low activity), and intermediate K yields coexistence of low- and high-activity regimes with strong basin separation.

Varying K reveals the resulting phase-transition structure:

Table 9. Phase transition characteristics by K

K range	Behavior	Separation degree
$K < 0.27$	Monostable (both converge to high activity)	≈ 0
$K \in [0.27, 0.70]$	Bistable (two-attractor multistability; hysteresis)	> 0.6
$K > 0.70$	Monostable (both converge to low activity)	≈ 0

In our scan, the maximum separation (0.675) occurs at $K = 0.557$, representing a balance where both attractors are well-formed.

5.5. Chat & Exercise to Express Induction

A key design feature of the present model is that Chat & Exercise provides information gain about interpersonal trust ($\text{IG}_W > 0$) while Express does not, creating an incentive for agents to first engage in curiosity-driven exploration before transitioning to coordination-driven participation. Empirical analysis of action transitions confirms this prediction: Chat & Exercise $\to$ Express transitions occur (164 instances across 10 seeds), while Express $\to$ Chat & Exercise transitions do *not* occur (0 instances), revealing a unidirectional induction pattern. Furthermore, periods of high Chat & Exercise activity are followed by modest increases in collective expression rate ($\text{Corr}(\bar{c}_t, \Delta\bar{e}_{t+1}) = +0.027$). To further validate this causal relationship, we applied Transfer Entropy analysis [43], which quantifies directed information flow between time series. The results show $\text{TE}(\text{Chat \& Ex.} \to \text{Express}) = 0.111$ bits versus $\text{TE}(\text{Express} \to \text{Chat \& Ex.}) = 0.043$ bits (ratio $= 2.57$, both $p < 0.001$), confirming significantly stronger information flow from Chat & Exercise to Express than vice versa. This supports the interpretation that Chat & Exercise serves as a "gateway" action enabling subsequent Express

 doi:10.20944/preprints202603.0933.v1

behavior, consistent with theatre workshop dynamics where participants first gauge group receptivity through low-stakes conversation and exercises before committing to expressive improvisation. Full details of the transition and Transfer Entropy analyses are provided in Supplementary Material (Sec. S-XI). A sensitivity analysis of the Express information gain assumption ($\mathrm{IG}_W(\text{Express}) = 0$) is provided in Supplementary Material (Sec. S-X).

5.6. Comparison with Simplified Model

To demonstrate the necessity of the key mechanisms in our model, we conducted a systematic chaotic itinerancy (CI) analysis comparing three model conditions:

1. **Full Model (Nonlinear)**: Default parameters with Hill function active ($c_{S,\text{hill}} = 0.5$, $K_{\text{hill}} = 0.4$).
2. **Linear Model**: Hill function coefficient set to zero ($c_{S,\text{hill}} = 0$), removing the nonlinear expression–trust coupling.
3. **No Coupling**: Expression–trust coupling strength set to zero ($K_{\text{hill}} = 0$), completely decoupling the expression dynamics from trust evolution.

Table 10 summarizes the CI detection results across 10 independent simulation runs with $N = 6$ agents over $T = 400$ time steps:

Table 10. Chaotic itinerancy detection across model conditions (10 seeds). CI Pass Rate indicates the fraction of runs exhibiting chaotic itinerancy (all 3 primary indicators positive).

Model Condition	CI Pass Rate	Mean Indicators	Seed=1 Result
Full Model (Nonlinear)	8/10 (80%)	2.40/3	PASS (3/3)
Linear ($c_{S,\text{hill}} = 0$)	2/10 (20%)	1.20/3	FAIL (2/3)
No Coupling ($K_{\text{hill}} = 0$)	0/10 (0%)	0.60/3	FAIL (1/3)

The results clearly demonstrate:

1. **Nonlinear Hill function enhances CI**: The full model achieves 80% CI pass rate compared to only 20% for the linear model—a fourfold improvement. The Hill function's sigmoidal response creates the threshold-crossing dynamics essential for attractor ruin formation.
2. **Expression–trust coupling is essential**: The no-coupling condition shows 0% CI pass rate with only 1/3 primary indicators consistently detected. Without the bidirectional feedback between expression and trust, the system converges to simple fixed-point dynamics rather than exhibiting chaotic itinerancy.
3. **Cluster structure differs qualitatively**: At seed=1, the full model detected 10 attractor ruins with 23% noise ratio (healthy transition states), while the linear model detected only 2 clusters with 56% noise ratio (excessive transitions), and the no-coupling condition showed 2 clusters with only 4% noise ratio (near-fixed-point behavior).

These findings establish that the nonlinear Hill function coupling is not merely a parameter choice but a **necessary structural feature** for generating the rich attractor landscape that supports chaotic itinerancy. The simplified linear model cannot produce the itinerant dynamics that characterize theatre workshop processes.

5.7. Parameter Sensitivity Analysis

To assess the robustness of our results to specific parameter choices, we conducted systematic sensitivity analyses for four key parameters: the Hill coefficient n, the half-saturation constant K, the precision modulation sharpness λ, and the trust–empowerment coupling strength $\kappa_{S\to u}$. Each parameter was tested for both bistability (multiple equilibria) and chaotic itinerancy (CI) across multiple seeds. Table 11 summarizes the main findings; full details are provided in Supplementary Material (Sec. S-VI).

Key findings: (i) *Hill coefficient* n: Bistability requires $n \geq 4$; CI robust at 70–85% for bistable values (default n=4: 85%). (ii) *Half-saturation* K: Bistable window $K \in [0.27, 0.70]$; CI 70–90% in $[0.20, 0.60]$ (default 0.40: 80%). (iii) *Precision* λ: CI robust (80–90%) across $[2, 16]$. (iv) *Coupling* $\kappa_{S\to u}$: Primary driver; CI increases monotonically 70%→100% (default 0.6: 80%).

 doi:10.20944/preprints202603.0933.v1

Table 11. Summary of parameter sensitivity analysis.

Parameter	Range	Bistable	CI Characteristics
Hill coeff. n	2–10	≥ 4	70–85% ($n \geq 4$); default 4 (85%)
Half-sat. K	0.2–0.8	0.27–0.70	70–90% in bistable; default 0.40 (80%)
Precision λ	2–16	All	Robust 80–90%; default 4.0
Coupling $\kappa_{S\to u}$	0.2–1.0	≥ 0.4	70%→100%; default 0.6 (80%)

In summary, the model's chaotic itinerancy is robust across a wide range of parameter values, with the trust–empowerment coupling strength acting as the primary control parameter. The Hill coefficient exhibits a critical threshold effect, while the precision modulation mechanism shows remarkable insensitivity to specific parameter choices.

6. Discussion

The numerical results establish chaotic itinerancy as the primary dynamical signature of our model: trajectories dwell near multiple quasi-stable collective modes and switch irregularly among them. This section discusses the theoretical implications of this finding, its relation to existing frameworks, and practical consequences for workshop facilitation.

6.1. On Social Priors and the Minimality of Preference Specification

Several Active Inference models of social behaviour introduce explicitly social or normative structure through prior preferences over outcomes. For example, Constant et al. formalize deontic value/cues, whereby observations can endow policies with normative salience [5], and related work on cooperative communication posits an adaptive prior belief of mental-state alignment [6]. More generally, Active Inference often represents goals/values as prior distributions over preferred observations [34,44].

While these approaches provide an alternative to utility functions, they raise a concern about how much of the social phenomenon is effectively specified in outcome-level preferences. When conformity or coordination is encoded directly in preference distributions, explanations risk circularity: the target pattern is assumed at the level of priors.

The present multi-agent EFE model adopts a different strategy. We do not attempt to remove priors from Active Inference, but we restrict *where* social structure is introduced: collective effects enter via state-transition structure and learning conditions, while preferences are restricted to broadly non-normative state variables (trust, empowerment, stamina) that shape feasibility and learning rather than directly prescribing a collective outcome.

Structural constraints vs. outcome-level priors: One might object that embedding Hill-type nonlinearity in state transitions constitutes a form of "structural prior knowledge" about social dynamics. We acknowledge this, but emphasize a crucial distinction. The Hill function specifies *affordances*—environmental conditions that make certain actions more or less feasible—rather than *preferences* over collective outcomes. Concretely, the Hill term encodes that (i) sustained expression is physiologically costly without group support, and (ii) trust builds more readily when collective participation exceeds a threshold. These are constraints on *what is possible* given the physical and psychosocial environment, analogous to how gravity constrains movement without prescribing where one should go. In contrast, outcome-level social priors would directly encode "high collective expression is preferred" or "conformity is good" in the preference distribution. Our model contains no such specification: agents prefer states of high trust, empowerment, and stamina for individual reasons, and collective patterns emerge from the interplay of these preferences with the affordance structure.

Moreover, empowerment preference updating is conditional on inferred trust. This yields *conditional preference plasticity* (a second-order dependence of learning on context) that can regulate exploration and learning without encoding a specific social goal at the outcome level. In a limited sense, the model involves a weak structural form of "social prior" insofar as preferences are defined

 doi:10.20944/preprints202603.0933.v1

over socially relevant latent variables; this differs from outcome-level priors that prescribe collective behaviour (e.g., conformity) as a preferred observation.

Accordingly, the collective phenomena in the simulations arise from individual inference, resource constraints, and coupling through shared latent variables and trust-gated learning, rather than from direct outcome-level social preference specifications.

6.2. Emergent Chaotic Itinerancy Without Social Priors

Our simulations exhibit chaotic itinerancy under EFE minimization without adding explicit social conformity utilities. Positive feedback loops are encoded in the *generative model* (state-transition dynamics) rather than as direct outcome-level preferences.

This result resonates with an early demonstration in the FEP literature: Friston et al. showed that an active-inference agent endowed with strong priors for Lorenz-attractor dynamics can generate chaotic trajectories through prediction-error suppression alone [13]. Our model extends this principle to a multi-agent social setting, where collective chaotic itinerancy emerges from individual EFE minimization without requiring explicit chaos-generating mechanisms in the policy.

Crucially, chaotic itinerancy differs from simple bistability. Bistable systems exhibit two well-separated attractors with deterministic convergence to one or the other. In contrast, CI involves *quasi-stable* regimes ("attractor ruins") that trajectories visit transiently before transitioning to other regimes. This structured variability—neither convergence to a single point nor random noise—captures the exploratory nature of real theatre workshop dynamics.

The three mechanisms play distinct roles:

1. **Cooperative cost reduction** creates immediate incentives to express when others do.
2. **Trust–empowerment coupling** amplifies empowerment gains in trusting environments.
3. **Comfort zone expansion** generates hysteresis by making preference changes irreversible.

6.3. Mechanisms of Collective Dynamics

The rich dynamical behaviors observed in the model—specifically chaotic itinerancy and the Chat & Exercise to Express induction—emerge from the interplay of three core mechanisms.

1. Trust–Empowerment Coupling (Global Amplification).

The coupling term $\kappa_{S \to u}$ creates a positive feedback loop between collective trust S and individual empowerment u. In high-trust regimes, the actions of others are perceived as more empowering, which in turn encourages further expression. This mechanism stabilizes the high-activity "attractor ruin," allowing the group to sustain periods of intense collaboration. Conversely, in low-trust states, this amplification is absent, trapping the group in quiescence until a fluctuation or intervention occurs.

2. Precision-Gated Preference Learning (Adaptive Hysteresis).

Instead of a fixed "comfort zone," agents dynamically update their preference parameters (μ_U) based on experience. Crucially, this learning is gated by trust: only when S is high do agents treat their empowerment gains as reliable signals for updating their self-model (increasing μ_U). This creates a "ratchet effect" where successful collective episodes leave a lasting trace in the agents' internal priors, preventing a simple return to the initial state and generating the path dependence observed in the intervention tests.

This mechanism aligns with psychological research on self-efficacy [28], which emphasizes that success experiences are most impactful in supportive environments. From a practical standpoint, this suggests that interventions should prioritize building trust before encouraging action. A temporary "push" toward high activity is ineffective unless the environment supports the internalization of positive experiences.

 doi:10.20944/preprints202603.0933.v1

3. Network-Mediated Induction (Local-to-Global Propagation).

A key innovation in the present model is the role of the social network and interpersonal trust W_{ij}. The "Chat & Exercise" action serves as a low-risk mechanism to build W_{ij} without requiring immediate global exposure. As local dyads build trust, the effective cost of "Express" decreases (via the synchrony term), eventually triggering a local cluster of expression. This local activity then feeds into the global trust S, potentially igniting a system-wide phase transition. This two-stage process—Chat & Exercise building local W_{ij}, leading to Express building global S—explains the unidirectional "Chat & Exercise $\rightarrow$ Express" induction observed in the results.

6.4. Functional Role of Model Mismatch

The agents' internal model intentionally omits the complex trust–empowerment coupling term. This "bounded rationality" design forces agents to treat the amplified empowerment from social interactions as positive *surprise* (prediction error) rather than a predicted outcome. This continuous stream of surprise drives exploration and prevents the system from settling into a static equilibrium, thereby sustaining the itinerant dynamics essential for creative collaboration. This mechanism complements the chaos-control perspective of [13]: whereas strong, accurate priors can *stabilize* chaotic trajectories via error suppression, our incomplete-model agents experience persistent prediction errors that *sustain* exploratory dynamics.

We investigated this question by comparing simulations under two conditions:

1. **Baseline (Incomplete Model)**: Agents predict empowerment without the trust coupling term (current implementation)
2. **Complete Model**: Agents include the full trust–empowerment coupling $\kappa_{S\rightarrow u}$ in their predictions

Behavioral results.

Surprisingly, the complete-model agents exhibit **zero collective expression**, collapsing entirely to the quiescent low-activity state (expression rate $\bar{e} = 0.00$, zero state transitions). Rather than exploiting their knowledge of the coupling to reach and maintain the high-activity attractor, omniscient agents become perfectly conservative. In contrast, the baseline (incomplete model) agents show robust chaotic itinerancy with an average of 55.2 ± 27.9 state transitions and expression rate $\bar{e} \approx 0.36$ per 400-step trajectory (Table 12).

Table 12. Comparison of Baseline vs. Complete Model agents ($N = 6$, $T = 400$, 5 runs).

Metric	Baseline (Incomplete)	Complete Model
Expression rate $\bar{e}$	0.364 ± 0.186	0.000 ± 0.000
State transitions	55.2 ± 27.9	0.0 ± 0.0
CI Diagnosis (Primary 3 Indicators, Section 5.1)		
1. Attractor ruins	8 clusters (**PASS**)	3 clusters (PASS*)
2. Residence time skewness	14.04 (**PASS**)	1.61 (PASS*)
3. FTLE signature	$\sigma = 0.128$ (**PASS**)	$\sigma = 0.027$ (**FAIL**)
CI Classification	**3/3 $\rightarrow$ CI = YES**	2/3 $\rightarrow$ CI = NO

CI diagnosis reveals spurious clustering.

Applying the three primary CI indicators (Section 5.1) to the complete-model trajectories yields an instructive pattern: indicators 1 (attractor ruins) and 2 (heavy-tailed residence times) pass, while indicator 3 (FTLE signature) fails due to insufficient variance ($\sigma = 0.027 < 0.05$). However, these "passing" indicators are **spurious**: they detect micro-fluctuations in the agents' *belief space* (the internal variables μ_S^i, μ_U^i that continue to update even without behavioral expression) rather than meaningful behavioral dynamics. The complete-model agents remain behaviorally frozen at zero expression throughout the entire simulation; the clustering algorithm merely partitions trivial belief-space wandering that has no behavioral consequence. This underscores the importance of interpreting CI indicators in conjunction with behavioral metrics: a trajectory with zero expression and zero state transitions cannot exhibit meaningful chaotic itinerancy regardless of belief-space clustering patterns.

 doi:10.20944/preprints202603.0933.v1

Interpretation.

This finding suggests that model mismatch is not merely a simplifying assumption but a **functional design feature** that promotes healthy collective dynamics:

- **Surprise as exploration driver**: Prediction errors generate surprise signals that prevent premature convergence
- **Avoiding coordination failure**: Complete knowledge of interdependencies can lead to "free-riding" expectations that precipitate collective inaction
- **Bounded rationality benefits**: Consistent with game-theoretic results showing that bounded rationality can improve collective outcomes [45]

This result aligns with ecological rationality perspectives arguing that cognitively simple heuristics can outperform optimal strategies in complex environments [46]. The agents' "ignorance" of the trust coupling generates the variability and exploration necessary for discovering and transitioning between collective modes—the very signature of chaotic itinerancy.

6.5. The Empowerment Expansion as Self-Recognition: Removal of Habitual Defenses

The latent empowerment expansion variable z_t^i (Eq. 31) deserves deeper theoretical interpretation. Rather than functioning as a mere computational switch, z_t^i represents the agent's **inference about whether it has crossed a threshold**—a form of *self-recognition* regarding one's own transformative state. In the framework of the Free Energy Principle, this can be understood as higher-order model selection: the agent infers "I am now in a growth mode" versus "this experience was circumstantial."

Connection to Theatre Studies: Breaking Habitual Defenses

This computational mechanism formalizes insights from theatre theory. Grotowski's concept of the "via negativa"—the removal of habitual psychological and physical defenses—describes precisely the kind of threshold-crossing that z_t^i captures [7]. The agent does not add new techniques but rather *removes* the protective shell that prevents authentic expression. When u_t^i exceeds $\mu_U^i + \theta_{\text{gap}}$, the agent recognizes that it has acted beyond its prior defensive boundary.

Accepting a New Self

The unidirectional nature of preference updates (the "ratchet effect") reflects a deeper psychological reality: once one has genuinely experienced expanded capability and recognized it as authentic ($z_t^i = 1$), the prior, more limited self-model becomes difficult to maintain. This is not mere behavioral change but a shift in *what the agent believes it can be*—an update to its generative model of self.

From the Active Inference perspective, z_t^i can be interpreted as evidence for a higher-level hypothesis: "I am someone who can act and express in this way." The precision-gated learning mechanism ensures that this evidence is only accepted when contextual conditions (trust) support reliable inference. In this sense, the model captures the interplay between environmental safety, threshold-crossing action, and identity transformation that practitioners of theatre have long recognized.

6.6. Quantitative Assessment of the Dual-Gate Mechanism

A natural question is whether the empowerment expansion gate z_t^i is *essential* for the ratchet effect, or whether trust-weighted precision alone suffices. We conducted a systematic comparison of three conditions (see Supplementary Material (Sec. S-VII) for full details):

1. **Double-Gate**: The current implementation, where preference updates require both high precision (trust-dependent) and mastery detection ($z_t^i = 1$)
2. **Single-Gate**: Precision weighting only, with mastery gate bypassed (always $\mathbb{E}[z_t^i] = 1$)
3. **No-Gate**: Both gates bypassed (always high precision, always updating)

The quantitative results (Table 13) reveal that the mastery gate provides **incremental improvement** rather than being strictly necessary for irreversibility. The ratchet ratio (upward/downward update events) was 4.5 for Double-Gate, 3.8 for Single-Gate, and 3.3 for No-Gate. All conditions exhibited

 doi:10.20944/preprints202603.0933.v1

net positive preference expansion (approximately 0.7 units), with similar maximum retracements (0.14–0.15 units).

Table 13. Comparison of preference update mechanisms across conditions (mean ± std, $n = 5$ runs).

Metric	Double-Gate	Single-Gate	No-Gate
Ratchet ratio	4.45 ± 1.26	3.82 ± 1.15	3.30 ± 0.77
Net change ($\Delta\mu_U$)	0.68 ± 0.54	0.75 ± 0.60	0.72 ± 0.57
Max retracement	0.15 ± 0.12	0.15 ± 0.12	0.14 ± 0.12
μ_U variance	0.08 ± 0.04	0.09 ± 0.05	0.08 ± 0.04

These results support the interpretation that the **primary source** of irreversibility is the unidirectional constraint (`upward_only=True`), which prevents downward preference updates regardless of gate configuration. The mastery gate z_t^i provides an **additional filtering mechanism** that:

- Increases the ratchet ratio by approximately 17% (from 3.8 to 4.5) compared to precision-only updates
- Filters learning signals to internalize only genuine threshold-crossing experiences
- Provides biological and psychological plausibility, as discussed in Section 6.5

The dual-gate design is therefore best understood as a *quality-enhancing* mechanism rather than an absolute requirement for irreversibility. It ensures that the learning signal is concentrated on pedagogically meaningful events—moments of genuine breakthrough—rather than being diluted by routine fluctuations. This is consistent with educational research emphasizing that not all successes are equally formative; breakthrough moments that exceed prior self-expectations carry special developmental significance.

Beyond preference learning dynamics, the gating mechanism operates within a dynamically rich substrate. Our chaotic itinerancy analysis (Section 5.6) demonstrates that the **nonlinear Hill function coupling is essential** for generating CI: the full model achieves 80% CI pass rate compared to only 20% for a linearized version and 0% when expression–trust coupling is disabled entirely. This establishes that the dual-gate mechanism filters transitions within a complex attractor landscape that itself depends critically on the nonlinear feedback structure.

6.7. EFE Component Analysis During Transitions

To validate the role of information gain in action selection, we analyzed the composition of Expected Free Energy (EFE) across simulation trajectories (see Supplementary Material (Sec. S-VIII) for full details).

The EFE decomposes into Risk and Information Gain (IG) components:

$$G(a) = \underbrace{\text{Risk}_S + \text{Risk}_U + \text{Risk}_H + \text{Risk}_W}_{\text{Risk terms}} - \underbrace{(\text{IG}_S + \text{IG}_u + \text{IG}_W)}_{\text{Information Gain terms}} \tag{59}$$

Analysis of the component breakdown reveals:

- **Risk composition**: Trust-related risk (Risk_S) dominates at 62%, followed by interpersonal trust (Risk_W, 29%) and empowerment (Risk_U, 6%).
- **IG composition**: Interpersonal trust information gain (IG_W), which is obtained *exclusively* through Chat & Exercise actions (Express and Rest yield $\text{IG}_W = 0$ by design), accounts for 44% of total IG across the simulation. This validates the design where Chat & Exercise provides unique epistemic value about social trust. Indirect empowerment gain (IG_u) contributes 37%, and direct trust observation (IG_S) contributes 20%.
- **Transition correlation**: EFE gradient magnitude ($|dG/dt|$) correlates positively with transition events ($r = 0.11$, $p < 0.01$), and IG shows positive correlation with transitions ($r = 0.31$, $p < 0.001$), suggesting that high information gain precedes or accompanies state transitions.

 doi:10.20944/preprints202603.0933.v1

These findings confirm that the Chat & Exercise action is selected not merely as a behavioral option but specifically for its epistemic value in reducing uncertainty about interpersonal trust—a core prediction of the Active Inference framework.

6.8. Relationship Between Chaotic Itinerancy and Mutual Entrainment

Our model exhibits a distinctive relationship between chaotic itinerancy (CI) and mutual entrainment (synchronization), which aligns with recent theoretical developments connecting these phenomena [47,48].

6.8.1. Kuramoto Order Parameter Analysis

To quantify the degree of mutual entrainment, we computed the Kuramoto Order Parameter (KOP) adapted for discrete action spaces (see Supplementary Material (Sec. S-IX) for details). The action synchrony is defined as:

$$R(t) = \left| \frac{1}{N} \sum_{j=1}^{N} e^{i\theta_j(t)} \right|, \quad \theta_j = \frac{2\pi a_j}{3} \tag{60}$$

where $a_j \in \{0, 1, 2\}$ denotes the action (Rest, Chat & Exercise, Express) of agent j. $R = 1$ indicates perfect synchronization (all agents choosing the same action), while $R \approx 0$ indicates uniform distribution across actions.

Figure 5 presents the KOP analysis under default parameters ($N = 6$, $T = 400$, seed $= 1$). The results reveal several key features:

6.8.2. Irregular Switching Between Synchronization and Desynchronization

The simulation exhibits **8 transitions** between synchronized ($R > 0.7$) and desynchronized ($R < 0.4$) states over 400 time steps, with a transition rate of 0.02 per step. This irregular switching is precisely the signature of CI as a mechanism for synchronization–desynchronization alternation described by Tsuda et al. [47].

Key statistics include:

- **Mean action synchrony**: $\bar{R} = 0.70$, SD $= 0.20$
- **Coefficient of variation**: CV$(R) = 0.29$ (high variability)
- **Synchronization ratio**: 45% of time in synchronized state
- **Mean synchronization duration**: 54 steps
- **Mean desynchronization duration**: 12 steps

The asymmetry between synchronization and desynchronization durations reflects the model's nonlinear dynamics: once agents achieve coordinated expression, the Hill function feedback sustains this state, whereas escape from desynchronization requires stochastic fluctuations or accumulated stamina recovery.

6.8.3. Theoretical Interpretation

This pattern supports two complementary perspectives on the CI–entrainment relationship:

CI as a Pathway to Intermittent Synchronization

Following Leyva et al. [48], CI can be viewed as a precursor to intermittent synchronization (IS). At intermediate coupling strengths, trajectories explore multiple quasi-stable modes before settling into synchronized episodes. The burst frequency (0.10) and intermittency index observed in our simulations indicate dynamics intermediate between pure CI and stable synchronization.

CI as a Mechanism for Synchronization–Desynchronization Switching

Following Tsuda et al. [47], CI provides a dynamical explanation for irregular alternation between coordinated and uncoordinated states. The high CV of R ($= 0.29$) and multiple transitions confirm that

 doi:10.20944/preprints202603.0933.v1

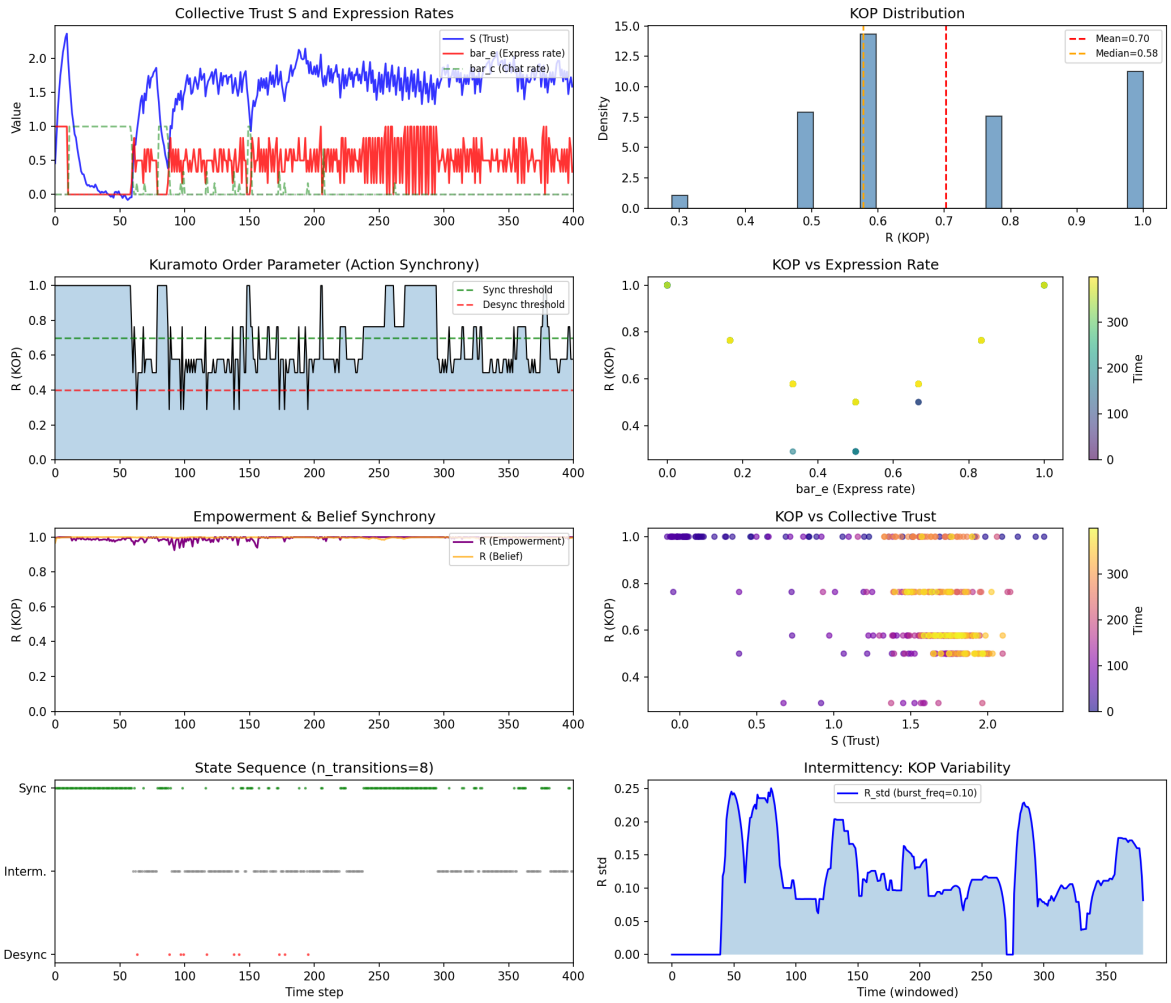

Figure 5. Kuramoto Order Parameter analysis under default parameters. (a) Time series of collective trust S and expression rate $\bar{e}$. (b) Action synchrony $R(t)$ with synchronization ($R > 0.7$) and desynchronization ($R < 0.4$) thresholds. (c) Empowerment and belief synchrony (both near $R = 1$). (d) State sequence showing synchronization–desynchronization transitions. (e) Distribution of R showing bimodal structure. (f–g) Relationships between KOP and collective variables. (h) Intermittency measure (local variability of R).

 doi:10.20944/preprints202603.0933.v1

the model does not settle into a single synchronization regime but actively explores both synchronized and desynchronized configurations.

6.8.4. Implications for Workshop Dynamics

This analysis suggests that workshop groups may naturally exhibit intermittent coordination:

- **Coordination is not all-or-nothing**: Groups alternate between periods of high synchrony (collective expression) and low synchrony (individual exploration or rest).
- **Desynchronization is functional**: Periods of low coordination may serve exploratory functions, allowing individuals to recover resources or explore alternative action modes.
- **Facilitator interventions**: The timing of facilitation may be most effective near transition points between synchronization and desynchronization states.

Importantly, empowerment synchrony ($R \approx 0.99$) and belief synchrony ($R \approx 0.998$) remain high throughout, indicating that action-level variability coexists with convergent internal states. This dissociation suggests that agents develop shared beliefs about the collective trust state while maintaining behavioral diversity—a pattern consistent with "thinking alike while acting differently."

6.9. Implications for Workshop Management

Priority of trust building Rather than immediately encouraging expression, it is important to first raise field trust. Comfort zone expansion does not occur unless trust exceeds the threshold.

Timing of intervention When trust is near the threshold, small interventions have large effects. Facilitators need to discern the "critical point."

Conditions for sustained change For "growth" accompanied by preference change rather than temporary excitement, successful experiences under sufficient trust are necessary.

6.10. Novelty Beyond Ising-Type Models

Ising-type social models typically specify an energy function (Hamiltonian) *a priori*; multistability (including bistability) and hysteresis follow as relaxation toward energy minima. The kinetic Ising model and social extensions [49,50] exemplify this setup, with Glauber dynamics yielding logit-type transition rates from the Hamiltonian.

Our model differs in five ways, reflecting a richer structural complexity:

(i) Action selection emerges from inference, not from an imposed energy function.

Although the update resembles a Glauber-type logit form, we do *not* posit a "social Hamiltonian" or "conformity utility." Action probabilities arise from Expected Free Energy (EFE) minimization, combining risk (KL divergence from preferred outcomes) and epistemic value. The decomposition $-\log P(e \mid h) = \text{Risk} + \text{Ambiguity}$ [4] provides a generative derivation rather than an *ad hoc* energy specification.

(ii) The latent field S is inferred, not externally imposed, and observation precision depends on action.

In standard Ising models, the external field is exogenous; here, collective trust S is a shared latent variable inferred from noisy observations via an Extended Kalman Filter. Crucially, observation noise $R(e)$ is action dependent: expressing ($e = 1$) yields more precise observations than remaining silent ($e = 0$). This "active observation" (action-modulated epistemic access) is central to POMDP/Active Inference architectures and not part of equilibrium Ising formulations.

(iii) Stamina H endogenizes effective temperature and action costs, with collective-state-dependent modulation.

The fatigue/resource constraint produces a time-varying effective inverse temperature β_{eff}, yielding intrinsic non-equilibrium dynamics. In addition, the Hill term $H_n(\bar{e}; K)$ reduces expression costs

 doi:10.20944/preprints202603.0933.v1

when collective activity is high, i.e., the group state modulates individual action costs. This feedback goes beyond the additive pairwise interactions typical of Ising extensions.

(iv) Comfort zone expansion generates learning-based path dependence distinct from equilibrium hysteresis.

Ising hysteresis under field sweeps is state-level and reversible in principle; here, path dependence is introduced by an irreversible preference-jump mechanism:

$$S > \theta_S \ \wedge \ \text{experience} > \mu_U \quad \Longrightarrow \quad \mu_U \leftarrow \mu_U + \Delta\mu \tag{61}$$

Even if the system returns to the same *state* $(S, \bar{e})$, the agent's internal preferences remain changed. Thus, history is encoded in parameter updates rather than in state trajectories, yielding a non-Markovian effect.

(v) Network topology and dyadic trust (W_{ij}) introduce local-global interplay.

Unlike mean-field Ising models where every agent interacts with the global average, our model incorporates a sparse social network where agents infer local interpersonal trust W_{ij}. These dyadic weights evolve based on interaction synchrony and are learned via Variational Message Passing, allowing agents to form distinct local relationships. This structure enables complex spatiotemporal patterns where local pockets of trust can emerge and propagate, creating a richer dynamical landscape than the uniform phase transitions of fully connected mean-field models. Furthermore, the specific epistemic role of Chat & Exercise in reducing uncertainty about W_{ij} adds a strategic dimension to local relationship building that is absent in standard spin systems.

In summary, the model can reproduce multistable collective dynamics but differs in generative structure: probabilistic inference under partial observability with action-dependent precision, endogenous non-equilibrium effects from resource constraints, learning-based path dependence via irreversible preference updates, and rich local-global network interactions.

Connection to Statistical Physics.

Despite these structural differences, the model exhibits characteristics familiar from statistical mechanics:

- **Phase transitions**: Sharp changes in collective behavior at critical parameter values.
- **Hysteresis**: The system's state depends on its history.
- **Critical slowing down**: Near the transition point, fluctuations are amplified.

The Hill function plays a role analogous to interaction energy in spin systems, with the Hill coefficient n controlling the sharpness of the transition, analogous to inverse temperature. In brief, our approach places the explanatory burden on inference under partial observability and learning rules in the generative model (including action-dependent observation precision and trust-gated preference plasticity), rather than on an externally specified social Hamiltonian.

6.11. Limitations and Future Directions

Several limitations of the present study also motivate concrete directions for future work.

6.11.1. Network Structure and Remaining Extensions

Our model extends beyond the mean-field approximation by embedding agents in an Erdős–Rényi network with dynamic interpersonal trust weights W_{ij}. Each agent observes a *local* expression rate $\bar{e}^{(i)}$ computed as a trust-weighted average over its neighborhood (Eq. 10), rather than the global expression rate $\bar{e}_t$. This network structure addresses several limitations of pure mean-field models:

 doi:10.20944/preprints202603.0933.v1

Advantages of the Network Extension

- **Heterogeneous influence**: Agents experience differentiated social influence based on their network position and interpersonal trust levels.
- **Spatial correlation**: Clustering and subgroup formation can emerge naturally from network topology.
- **Appropriate for small groups**: Sparse networks with $k_{\text{avg}} = 4$ are more realistic for typical workshop sizes of 8–20 participants.
- **Curiosity-driven exploration**: The Chat & Exercise action provides information gain about interpersonal trust, implementing a distinct behavioral mode from expression-driven coordination.

Remaining Directions

While the network extension addresses the core limitation of mean-field coupling, several research directions remain:

1. **Topology effects**: How do small-world, scale-free, or clustered network structures affect chaotic itinerancy? The current Erdős–Rényi topology could be replaced with more structured alternatives [2].
2. **Dynamic networks**: The current model uses fixed network topology with dynamic edge weights. Extending to time-varying network structure (edge creation/deletion) would capture evolving interaction patterns.
3. **Role differentiation**: Workshop facilitators could be modeled as hub nodes with enhanced connectivity or modified action preferences.

6.11.2. Agent Heterogeneity and Further Action Richness

Our model extends the action space from binary (rest/express) to ternary (rest/chat & exercise/express), enabling distinct behavioral modes for curiosity-driven social exploration (Chat & Exercise) versus expression-driven coordination (Express). However, agents remain homogeneous in their parameters and preferences. Future extensions could introduce:

- **Heterogeneous agents**: Variation in baseline stamina, trust sensitivity, or preference parameters across agents.
- **Graded actions**: Continuous-valued expression intensity rather than discrete actions.
- **Role-specific preferences**: Facilitators with different preference structures or action costs.

6.11.3. Empirical Grounding

A promising next step is to operationalize the model's notion of "expression" and map it to observable features extracted from real-time workshop recordings. In practice, expressive acts can be represented by time-stamped behavioral markers derived from video and audio streams, such as speaking turns, gesture amplitude, body orientation toward the group, movement initiation, prosodic intensity, interpersonal distance changes, and dyadic synchrony measures. These features can be aggregated into a continuous-valued proxy for expression intensity, thereby replacing the binary action variable with a measured or inferred action process.

Given such observations, the model can be extended into a state-space estimation problem in which latent variables—most notably the field-level trust state S_t and participant-level empowerment u_t^i (and potentially fatigue-related states beyond the present stamina proxy H_t^i)—are inferred from multimodal data. This can be approached using standard Bayesian filtering and smoothing techniques (e.g., extended/unscented Kalman filtering, particle filtering, or variational message passing) applied to a suitably parameterized observation model that links latent states to behavioral features. In this empirical extension, workshop facilitation becomes a partially observed control problem: interventions can be treated as exogenous inputs that transiently perturb action tendencies or observation precision, allowing one to quantify how timing and intensity of facilitation shift the system across critical thresholds.

 doi:10.20944/preprints202603.0933.v1

Beyond retrospective analysis, a longer-term goal is to enable real-time, model-based support for workshop facilitation. If latent trust and empowerment states can be tracked online, the model may provide principled indicators of proximity to transition points (critical slowing down, rising variance, or increasing sensitivity to perturbations), thereby informing when minimal interventions are most effective. Such an approach would move toward an adaptive facilitation loop, in which the workshop is managed not by enforcing predetermined outcomes, but by monitoring and supporting the conditions under which constructive collective transitions—such as sustained engagement and comfort-zone expansion—can emerge.

Finally, developing empirical protocols will require careful attention to ethics, privacy, and interpretability. Any real-time inference system must respect participants' consent and data governance, and should be designed to augment, rather than replace, human facilitation. Nevertheless, the present model suggests a principled pathway from theory to practice: by linking expressive behavior to observable features and treating workshop dynamics as a coupled inference-and-control process, Active Inference-based models may contribute to richer, evidence-informed workshop design and facilitation.

7. Conclusions

This study has developed a multi-agent model of collective dynamics in theatre workshops, grounded in Active Inference and Expected Free Energy minimization. The central finding is that **chaotic itinerancy**—structured switching among multiple metastable collective modes—emerges naturally from EFE minimization with trust-gated preference learning, without prescribing collective outcomes as preferred observations.

The main contributions are organized into two categories: primary theoretical contributions and the methodological elements that support them.

Primary Contributions

1. **Emergent Chaotic Itinerancy Without Outcome-Level Social Priors**: We demonstrated that chaotic itinerancy can emerge from EFE minimization when collective effects are encoded in the *predictive model* (state transition dynamics) rather than in explicit social preference terms. Trajectories dwell near multiple quasi-stable "attractor ruins" and switch irregularly among them, producing structured variability rather than convergence to a single fixed point. This maintains the principled derivation of behavior from probabilistic inference while avoiding the circularity of directly prescribing collective outcomes through priors. Crucially, this result successfully reproduces the dynamic process of theatre workshops—where groups alternate between periods of high energy and quiescence—providing a novel computational approach to theatre analysis that has not been seen in previous studies. Supplementary checks confirmed intervention responsiveness and the presence of multiple coexisting metastable regimes.
2. **Precision-Gated Preference Learning**: Rather than employing *ad hoc* threshold rules for preference changes, we derived a principled variational mechanism in which preference parameters are hierarchically embedded as latent hyperstates subject to precision-weighted Bayesian updating. Trust modulates learning precision, providing a principled foundation for "comfort zone expansion" that generalizes previous threshold-based formulations.
   In the Theatre workshop interpretation, this hierarchical mechanism captures a clinically salient "threshold-crossing" process. Under sufficient inferred trust, positive empowerment overshoots are internalized as durable preference shifts, yielding a qualitative change in an agent's expressive agency. This provides a minimal computational account of the transition from inhibited participation to sustained risk-taking self-disclosure without prescribing the collective outcome at the level of social priors.

Methodological and Technical Contributions

To realize and validate the above contributions, we developed the following technical framework:

 doi:10.20944/preprints202603.0933.v1

3. **Interpersonal Trust Network with Local Averaging**: We extended the mean-field collective dynamics to an explicit network model, where agents are embedded in a sparse Erdős–Rényi graph with dynamic interpersonal trust weights W_{ij}. Local expression rates are computed as trust-weighted averages over each agent's neighborhood, avoiding complete mean-field approximation while maintaining computational tractability through variational message passing with Jaakkola bounds. Information gain about interpersonal trust is obtained exclusively through Chat & Exercise actions, implementing curiosity-driven social exploration as a distinct behavioral mode.
4. **Robust Parameter Regime for Chaotic Itinerancy**: Systematic parameter scans demonstrate that chaotic itinerancy is robust across a wide range of Hill coefficients, coupling parameters, and noise levels. This establishes that the model's qualitative behavior is not critically dependent on specific parameter choices.
5. **Reproducible Verification Protocols**: We provided complete simulation protocols, parameter tables, and verification procedures in the appendices, with chaotic-itinerancy diagnostics as the primary test supplemented by intervention response and multiple-equilibria checks.

The key theoretical insight is that structured variability in collective social systems can be understood through the principled lens of Active Inference. By embedding Hill-type nonlinearities, trust–empowerment coupling, and precision-gated learning within agents' generative models, we obtain chaotic itinerancy as an emergent property of EFE minimization—without recourse to outcome-level social priors that would directly encode the collective patterns to be explained.

From a practical standpoint, the model suggests that workshop facilitation shapes not which single "attractor" the group reaches, but rather the *itinerant route* by which the group explores the space of collective possibilities. Interventions can redirect trajectories among metastable modes, but the system's inherent variability means that outcomes are shaped by ongoing interaction rather than determined by initial conditions alone.

The chaotic-itinerancy analysis—our primary numerical contribution—demonstrates that workshop-like dynamics naturally exhibit metastable switching among multiple collective modes. Across a broad parameter set, trajectories exhibit the hallmarks of CI: multiple attractor ruins, irregular switching, and structured variability consistent with the "attractor ruin" picture [10–12,36]. This supports an interpretation in which workshops alternate among transient modes (inhibited, exploratory, playful, confrontational) before settling, and in which facilitation may shape the sequence of visited modes rather than determining a single outcome.

Future directions include exploring alternative network topologies (small-world, scale-free), incorporating agent heterogeneity, enabling dynamic network structure, and developing empirTical protocols for tracking latent states from behavioral data in real workshop settings.

Author Contributions: Conceptualization, S.M. and T.S.; Methodology, S.M. and T.S.; Formal Analysis, T.S.; Investigation, S.M.;Writing—Original Draft Preparation, S.M. and T.S.; Writing—Review and Editing, S.M.; Software, T.S.; Visualization, T.S.; Validation, S.M.; Supervision, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. All numerical results are based on simulations and can be reproduced using the model specification, parameter tables (Supplementary Material (Sec. S-I)), and simulation algorithm (Supplementary Material (Sec. S-XIV)) provided herein. The reference implementation can be shared for peer review as an anonymized archive upon request; a public repository link will be provided upon publication. All simulations were implemented in Python. The simulation code and figure-generation scripts can be made available to reviewers upon request, and may be shared as an anonymised archive if required. Parameter values and a step-by-step simulation algorithm are provided in the appendices to support independent reproduction. A public repository link will be provided upon publication.

 doi:10.20944/preprints202603.0933.v1

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol.* **1978**, *83*, 1420–1443. https://doi.org/10.1086/226707.
2. Castellano, C.; Fortunato, S.; Loreto, V. Statistical Physics of Social Dynamics. *Rev. Mod. Phys.* **2009**, *81*, 591–646.
3. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138.
4. Da Costa, L.; Tenka, S.; Zhao, D.; Sajid, N. Active inference as a model of agency. *arXiv [cs.AI]* **2024**.
5. Constant, A.; Ramstead, M.J.D.; Veissiere, S.P.L.; Friston, K. Regimes of expectations: An active inference model of social conformity and human decision making. *Front. Psychol.* **2019**, *10*, 679.
6. Vasil, J.; Badcock, P.B.; Constant, A.; Friston, K.; Ramstead, M.J.D. A world unto itself: Human communication as active inference. *Front. Psychol.* **2020**, *11*, 417.
7. Grotowski, J. *Towards a Poor Theatre*; Methuen: New York, NY, 1968.
8. Brook, P. *The Empty Space*; Penguin Books: London, England, 1968.
9. Boal, A. *Theatre of the Oppressed*; Pluto Press: London, England, 1979.
10. Kaneko, K. Clustering, coding, switching, hierarchical ordering, and control in a network of chaotic elements. *Physica D* **1990**, *41*, 137–172.
11. Tsuda, I. Chaotic itinerancy as a dynamical basis of hermeneutics in brain and mind. *World Futures* **1991**, *32*, 167–184.
12. Tsuda, I. Chaotic itinerancy. *Scholarpedia J.* **2013**, *8*, 4459.
13. Friston, K.J.; Daunizeau, J.; Kiebel, S.J. Reinforcement learning or active inference? *PLoS One* **2009**, *4*, e6421.
14. Lauro Grotto, R.; Guazzini, A.; Bagnoli, F. Metastable structures and size effects in small group dynamics. *Front. Psychol.* **2014**, *5*, 699.
15. Liebovitch, L.S.; Peluso, P.R.; Norman, M.D.; Su, J.; Gottman, J.M. Mathematical model of the dynamics of psychotherapy. *Cogn. Neurodyn.* **2011**, *5*, 265–275.
16. Schiepek, G.K.; Tominschek, I.; Heinzel, S. Self-organization in psychotherapy: testing the synergetic model of change processes. *Front. Psychol.* **2014**, *5*, 1089.
17. Heinzel, S.; Tominschek, I.; Schiepek, G. Dynamic patterns in psychotherapy–discontinuous changes and critical instabilities during the treatment of obsessive compulsive disorder. *Nonlinear Dynamics Psychol. Life Sci.* **2014**, *18*, 155–176.
18. Schiepek, G.; Stutzle, R.; Aichhorn, W. The mathematics of psychotherapy: A synergetic model of change dynamics. *Nonlinear Dynamics Psychol. Life Sci.* **2016**, *20*.
19. Magerko, B.; Manzoul, W.; Riedl, M.; Baumer, A.; Fuller, D.; Luther, K.; Pearce, C. An empirical study of cognition and theatrical improvisation. In Proceedings of the Proceedings of the seventh ACM conference on Creativity and cognition, New York, NY, USA, 2009. https://doi.org/10.1145/1640233.1640253.
20. Murphy, M. Imaginative play for a predictive spectator: theatre, affordance spaces, and predictive engagement. *Phenomenol. Cogn. Sci.* **2022**, *21*, 1069–1088.
21. Albarracin, M.; Constant, A.; Friston, K.J.; Ramstead, M.J.D. A variational approach to scripts. *Front. Psychol.* **2021**, *12*, 585493.
22. Hill, A.; Hill, A.; Paganini-Hill, A. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *The Journal of Physiology* **1910**, *40*, 4–7.
23. Conger, J.A.; Kanungo, R.N. The empowerment process: Integrating theory and practice. *Acad. Manage. Rev.* **1988**, *13*, 471–482.
24. Bandura, A. Self-efficacy mechanism in human agency. *Am. Psychol.* **1982**, *37*, 122–147.
25. Klyubin, A.S.; Polani, D.; Nehaniv, C.L. Empowerment: A universal agent-centric measure of control. In Proceedings of the 2005 IEEE Congress on Evolutionary Computation. IEEE, 2005, Vol. 1, pp. 128–135.
26. Edmondson, A. Psychological safety and learning behavior in work teams. *Adm. Sci. Q.* **1999**, *44*, 350–383.
27. Edmondson, A.C.; Lei, Z. Psychological safety: The history, renaissance, and future of an interpersonal construct. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2014**, *1*, 23–43.
28. Bandura, A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.* **1977**, *84*, 191–215.
29. Ryan, R.M.; Deci, E.L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **2000**, *55*, 68–78.
30. Zajonc, R.B. Social facilitation: A solution is suggested for an old unresolved social psychological problem. *Science* **1965**, *149*, 269–274.

 doi:10.20944/preprints202603.0933.v1

31. Borg, G.A. Psychophysical bases of perceived exertion. *Med. Sci. Sports Exerc.* **1982**, *14*, 377–381.
32. Centola, D.; Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **2007**, *113*, 702–734.
33. Joyner, M.J.; Coyle, E.F. Endurance exercise performance: the physiology of champions: Factors that make champions. *J. Physiol.* **2008**, *586*, 35–44.
34. Parr, T.; Pezzulo, G. Understanding, explanation, and active inference. *Front. Syst. Neurosci.* **2021**, *15*, 772641.
35. Simon, H.A.; Others. Theories of bounded rationality. *Decision and organization* **1972**, *1*, 161–176.
36. Tsuda, I. Dynamic link of memory—Chaotic memory map in nonequilibrium neural networks. *Neural Netw.* **1992**, *5*, 313–326.
37. Mierski, N.; Pilarczyk, P. Analysis of the chaotic itinerancy phenomenon using entropy and clustering. *arXiv [nlin.CD]* **2025**.
38. Namikawa, J. Chaotic itinerancy and power-law residence time distribution in stochastic dynamical system. *arXiv [nlin.CD]* **2004**. Published as Phys. Rev. E 72, 026204 (2005)., https://doi.org/10.1103/PhysRevE.72.026204.
39. Sauer, T. Chaotic itinerancy based on attractors of one-dimensional maps. *Chaos* **2003**, *13*, 947–952.
40. Tsuda, I.; Umemura, T. Chaotic itinerancy generated by coupling of Milnor attractors. *Chaos* **2003**, *13*, 937–946.
41. Fujimoto, K.; Kaneko, K. Bifurcation cascade as chaotic itinerancy with multiple time scales. *Chaos* **2003**, *13*, 1041–1056.
42. Menck, P.J.; Heitzig, J.; Marwan, N.; Kurths, J. How basin stability complements the linear-stability paradigm. *Nat. Phys.* **2013**, *9*, 89–92.
43. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
44. Torresan, F.; Kanai, R.; Baltieri, M. Prior preferences in active inference agents: soft, hard, and goal shaping. *arXiv [cs.AI]* **2025**.
45. Gigerenzer, G.; Selten, R. *Bounded rationality: The adaptive toolbox*; MIT press: London, England, 2002.
46. Todd, P.M.; Gigerenzer, G. *Ecological rationality: Intelligence in the world*; Evolution and Cognition, Oxford University Press: Cary, NC, 2012.
47. Tsuda, I.; Fujii, H.; Tadokoro, S.; Yasuoka, T.; Yamaguti, Y. Chaotic itinerancy as a mechanism of irregular changes between synchronization and desynchronization in a neural network. *J. Integr. Neurosci.* **2004**, *3*, 159–182.
48. Leyva, I.; Sendina-Nadal, I.; Letellier, C.; Sevilla-Escoboza, J.R.; Vera-Aila, V.P. From chaotic itinerancy to intermittent synchronization in complex networks. *arXiv [nlin.AO]* **2025**.
49. Brock, W.A.; Durlauf, S.N. Discrete Choice with Social Interactions. *Rev. Econ. Stud.* **2001**, *68*, 235–260.
50. Durlauf, S.N. How can statistical mechanics contribute to social science? *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 10582–10584.