BELIEFS IN FEEDBACK LOOPS: FACTORED REPRESENTATIONS OF INTERACTING AGENTS IN TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex systems, from ecosystems to brains, are usually modeled as modular components coupled through feedback rather than as indivisible wholes. Can a transformer trained on the joint behavior of interacting agents be understood in the same way? In this work, we develop a computational-mechanics framework to study agents interacting in feedback loops. We formally prove that any joint stochastic process can be presented as resulting from the behavior of agents interacting in feedback, and the resulting process is generated by a hidden Markov model whose Bayesian beliefs factorize into separate components per agent. Furthermore, our experiments show that transformers trained on such feedback sequences make use of this separability, disentangling the beliefs about the underlying factors into separate subspaces of their residual stream. This representation is found to scale linearly (instead of exponentially) with the number of agents. Interventions show that these representations are causally efficacious, allowing for targeted steering of individual agents' beliefs by selectively inserting counterfactuals. However, coarse-graining the interface is found to disrupt the factorization, and can force a higher-dimensional representation despite less information. Together, the results establish both the dimensional advantage of a factorized representation and transformers' ability to exploit it, laying the groundwork for investigating traces of such compositional predictive architecture in pretrained models.

1 INTRODUCTION

Communities can accomplish feats beyond the capacity of their individuals—groups of ants solve complex geometry problems (Dreyer et al., 2025), birds flock in murmurations responding to predators as a single perceptive body (Cavagna et al., 2010; Attanasi et al., 2014), AI agents coordinate to effectively infiltrate digital infrastructure (Greenblatt et al., 2026), and people collaborate in the global project of developing a scientific understanding of the universe (Fortunato et al., 2018; Zollman, 2007; Hull, 1988). Interestingly, the problem of understanding how individuals come together into collectives is dual to the question of how a single complex system can be understood as a composite of sub-components. A longstanding tradition in cognitive science interprets single cognitive systems as arising from the interaction of distinct modular 'agents' (Minsky, 1986). Work in multiagent systems further seeks to formalize interactions among multiple agents, and perhaps more importantly, how agents form beliefs that can motivate predictions about other agents (Gmytrasiewicz & Doshi, 2005; Albrecht & Stone, 2018; Chen et al., 2020; Zhang et al., 2022). Similarly, there is a substantial body of work in mechanistic interpretability aiming to decompose deep learning systems into interconnected 'circuits' (Olah, 2022; Olsson et al., 2022; Wang et al., 2022).

The above considerations raise a question for any observer of a complex system: under what conditions can the observer track the system not as an indivisible whole, but via tracking its sub-components separately and then putting these together?

As a first step towards addressing this question, here we investigate transformers trained to predict data generated by interacting agents. Following computational mechanics (Crutchfield, 2012), we describe the behavior of agents using *transducers* (Barnett & Crutchfield, 2015)—stochastic automata that generate outputs and update their memory via a probabilistic rule. Recent work has used principles of computational mechanics to show that transformers tend to represent the geometry of Bayesian beliefs within their internal activations during inference (Shai et al., 2024; Piotrowski et al., 2025; Shai et al., 2026). Building on these works, we investigate whether a transformer trained on data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generated by interacting agents builds internal representations that mirror the agents that generated the data.

In this paper we introduce a theoretical framework to formally investigate the prediction of interacting agents, which is then empirically validated on trained transformers. The theoretical contributions are:

1. We develop a general computational mechanics framework to describe the behavior of agents arranged in feedback loops.
2. We formally determine the conditions under which a joint process can be decomposed as resulting from interacting subcomponents.
3. We demonstrate that these factored processes—if sufficiently small—lead to exponential advantages in the embedding dimension.

We use this theory to derive the following experimental results on transformers trained to predict the activity of interacting agents:

4. Transformers are found to linearly represent factored Bayesian beliefs, where each factor tracks individual agents.
5. These representations are causally efficacious, being separately steerable by inserting counterfactual beliefs on specific sub-spaces.
6. By using factored beliefs, transformers can attain exponential savings in the required dimensionality for effectively predicting interacting agents. However, these savings can be broken by coarse-graining.

2 Agents in Feedback Loops

2.1 Background

In this work we instantiate the behavior of agents using *transducers* (Barnett & Crutchfield, 2015): stochastic automata that generate outputs $Y_t \in \mathcal{Y}$ from receiving inputs $X_t \in \mathcal{X}$ using a memory state $R_t \in \mathcal{R}$.[1] A transducer $\mathbf{T} = \langle \mathcal{X}, \mathcal{Y}, \mathcal{R}, T^{(y|x)}_{r\to r'}, p(r_0)\rangle$ is specified via the Markov kernel

$$T^{(y|x)}_{r\to r'} \equiv \Pr(Y_t = y, R_{t+1} = r' | X_t = x, R_t = r), \tag{1}$$

together with an initial state distribution $p(r_0) = \Pr(R_0 = r_0)$. The probability of a transducer generating a given input-output sequence $(x_{0:L}, y_{0:L})$ can be calculated as

$$p(y_{0:L}|x_{0:L}) = \sum_{r_{0:L+1}} p(r_0) \prod_{t=0}^{L-1} T^{(y_t|x_t)}_{r_t\to r_{t+1}}. \tag{2}$$

Thus, the effective behavior of an agent—its input-output distribution, as in Eq. (2)—can be captured via an appropriate transducer. In computational mechanics jargon, we say the transducer provides a *presentation* of the behavior of that agent (Rosas et al., 2025).

It is convenient to describe transducers via linear operators of the form

$$\hat{T}^{(y|x)} \equiv \sum_{r,r'} \boldsymbol{e}_{r'} T^{(y|x)}_{r\to r'} \boldsymbol{e}_r^\intercal, \tag{3}$$

where $\boldsymbol{e}_r$ is a $|\mathcal{R}|$-dimensional vector with 1 in the r-th place and zero everywhere else. This operator can be used to build the *Mixed State Presentation* (MSP) of a transducer, which corresponds to vectors of probabilities that are updated by Bayesian filtering via

$$\boldsymbol{\eta}_t \propto \prod_{i=0}^{t-1} \hat{T}^{(y_i|x_i)} \boldsymbol{\eta}_0, \tag{4}$$

with $\boldsymbol{\eta}_0 \in \Delta(\mathcal{R})$ being a prior distribution. The belief state $\boldsymbol{\eta}_t \in \Delta(\mathcal{R})$ is the posterior distribution of R_t given the history $(x_{0:t}, y_{0:t})$, and can be shown to be a sufficient statistic of the past regarding future events. Thus, the geometry of belief states implies that proximity in the simplex guarantees similar futures. Belief states have been observed in the residual stream of LLMs trained to predict a process (Shai et al., 2024; Piotrowski et al., 2025). In what follows, we show that the belief state of a feedback loop between agents contains the belief states of each agent contained within the loop.

[1] Uppercase letters (e.g. X, Y) are used to denote random variables and lowercase (e.g. x, y) their realizations, and calligraphic letters (e.g. $\mathcal{X}, \mathcal{Y}$) denote the sets over which they take values. We also use the following abbreviations: $x_{a:b} = (x_a, \ldots, x_b)$, $x_{:b} = x_{0:b}$, $x_{a:} = x_{a:\infty}$, and $x_{:} = x_{0:\infty}$.

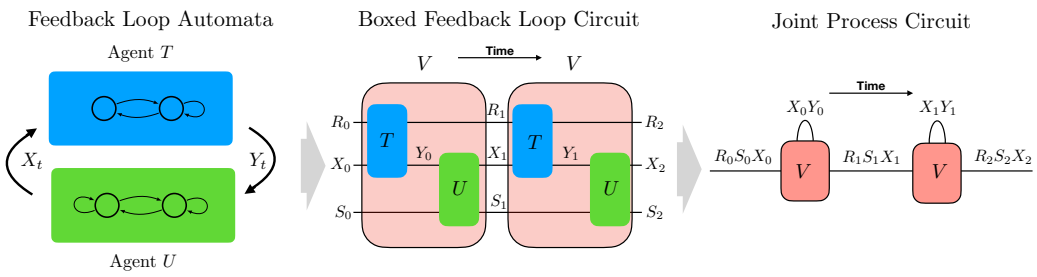


Figure 1: (**Left**) Two automata for agent T and agent U interact in a feedback loop. (**Center**) The feedback loop circuit specifies the time-ordering of agent T and agent U in a feedback loop. The feedback circuit can be boxed into a single repeating circuit element V. (**Right**) The feedback loop can be expressed as an HMM for the joint process.

2.2 Computational Mechanics of Transducers in Feedback Loops

Agents can be linked into perception-action feedback loops, either with an environment or with each other (Tishby & Polani, 2011; Fiderer et al., 2025; Boyd et al., 2025; Baltieri et al., 2026). In a cycle between two agents T and U, the actions Y of agent T become agent U's perceptions, and agent U's actions X become agent T's perceptions (see Fig. 1, left). While this picture is intuitive, it leaves ambiguity in the time index of the cycle. To index a time series between the agents, this symmetry needs to be broken by (w.l.o.g.) setting T to operate first (see Fig. 1, center). The result is a joint process over the perceptions and actions together with the latent states R_t of agent T and S_t for agent U (see Fig. 1, right).

The observable joint process arises by marginalizing over these latent states:

$$p(x_{0:L}, y_{0:L}) = \sum_{\substack{x_L \\ s_{0:L+1} \\ r_{0:L+1}}} p(r_0, s_0, x_0) \prod_{t=0}^{L-1} T^{(y_t|x_t)}_{r_t \to r_{t+1}} U^{(x_{t+1}|y_t)}_{s_t \to s_{t+1}}, \tag{5}$$

where $p(a) = \Pr(A = a)$ is a short-hand notation. See Appendix D for more technical details.

Under what conditions can a bivariate time-series process be interpreted as resulting from the interactions of two agents entangled in feedback? Our first result leverages the chain rule for causal conditioning (Massey, 1990) to show that this can always be done (proof in Appendix G).

Theorem 1 (Feedback Universality)**.** *For any joint process $p(x_{0:L}, y_{0:L})$, there exist transducers T and U such that the feedback loop between them produces that process.*

In essence, in analyzing the dynamics of feedback processes, our results apply to any joint process through decomposition. This framework goes meaningfully beyond feed-forward dynamics described in Appendix B and Boyd et al. (2025).

2.3 A Holistic Model for Interacting Agents

Our next result shows that, despite the intricacies of feedback between transducers, their joint process can always be modeled via a single HMM V whose linear operator is a tensor including two components—one for agent U and one for agent T.

Theorem 2. *The linear operator for a valid joint HMM of the feedback process between agent T and agent U is the tensor product*

$$\hat{V}^{(xy)} = \hat{T}^{(y|x)} \otimes \hat{\mathcal{U}}^{(x|y)}, \quad \textit{where} \quad \hat{\mathcal{U}}^{(x|y)} \equiv \sum_{z' \in \mathcal{X}} \hat{U}^{(z'|y)} \otimes \boldsymbol{e}_{z'} \boldsymbol{e}_x^{\intercal} \tag{6}$$

is the action-delayed linear operator for agent U.

Proof. See Appendix D. □

Note that there are other valid presentations for the feedback loop, but we will focus on this particular presentation that breaks the symmetry between agent T and agent U by assigning T to operate first in producing X_t and Y_t simultaneously (see Appendix D).

The joint feedback HMM V arises, as shown in Fig. 1, by "boxing" the circuit elements of agent U and T into a single element V. Thus, the linear operators $\hat{V}^{(xy)}$ produce the joint feedback process according to standard HMM rules $p(x_{0:L}, y_{0:L}) = \mathbf{1}^\intercal \prod_{t=0}^{L-1} \hat{V}^{(x_t y_t)} \boldsymbol{\eta}_0^{R,S,X}$, where $\boldsymbol{\eta}_0^{R,S,X} \equiv \sum_{r_0,s_0,x_0} (\boldsymbol{e}_{r_0} \otimes \boldsymbol{e}_{s_0} \otimes \boldsymbol{e}_{x_0})\, p(r_0, s_0, x_0)$ is the initial probability vector in the feedback process's memory space and $\mathbf{1} \equiv \sum_{r_0,s_0,x_0} (\boldsymbol{e}_{r_0} \otimes \boldsymbol{e}_{s_0} \otimes \boldsymbol{e}_{x_0})$ is the vector of all 1's in the same space. The new memory space is the tensor product space $\mathcal{R} \otimes \mathcal{S} \otimes \mathcal{X}$, which leads to an increase in the apparent dimension of the system as a whole.

Corollary 1. *The Mixed State Presentation of the joint feedback HMM V is in a $(|\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1)$-dimensional simplex.*

Proof. $\hat{V}$ operates on $\mathcal{R} \otimes \mathcal{S} \otimes \mathcal{X}$. The simplex of states in this space has $|\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1$ dimensions. □

This result hints at the complexity of the feedback process. While there are many definitions of complexity (Li & Vitányi, 2008; Ruebeck et al., 2018; Crutchfield, 2012; Gu et al., 2012), a practically relevant measure in the context of machine learning is the number of latent dimensions necessary to predict an interface. This is lower-bounded by the fractal dimension of the mixed state presentation (Jurgens & Crutchfield, 2021b), which can be hard to calculate in general (Jurgens & Crutchfield, 2021a). Conversely, the minimum dimension is upper-bounded by the simplex dimension of a presentation that can produce the desired process. For independent and conditionally dependent processes without feedback connections, it has been shown that this is precisely the dimension of the latent space LLMs use to make predictions when trained on next-token prediction (Shai et al., 2024; Piotrowski et al., 2025).

Addressing the complexity of feedback processes, the explicit construction of the joint HMM indicates that at most $|\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1$ dimensions are required to predict. This first cut at complexity indicates that agents interacting multiply each other's dimensional complexity. However, as we will show, this depends on how factorizable these agents are.

3 Multi-Agent Belief Geometry

Having established a general framework for describing feedback processes between agents, we next show how the belief update in joint processes can be interpreted as a product of the individual agents' updates, and demonstrate that transformer networks linearly represent and causally rely on this representation in their residual stream. Building on the factored world hypothesis (Shai et al., 2026), we distill these notions into the following testable hypothesis:

- **Multi-Agent Belief Geometry Hypothesis (MABGH):** The belief state geometry of each agent within a feedback loop appears within the latent activations of a transformer neural network trained on next-token prediction of the joint feedback process and is causally load-bearing.

To test this hypothesis, we first prove that if the belief states of two agents are expressible as a tensor product, this factorization remains under belief updating (proof in Appendix E):

Corollary 2. *The belief states of the joint feedback process between the agent T and the agent U remain in a tensor product between T's memory space $\mathcal{R}$ and the joint space of the agent U's memory and action $\mathcal{S} \otimes \mathcal{X}$: $\boldsymbol{\eta}_t^{R,S,X} = \boldsymbol{\eta}_t^R \otimes \boldsymbol{\eta}_t^{S,X} \;\Rightarrow\; \boldsymbol{\eta}_{t+1}^{R,S,X} = \boldsymbol{\eta}_{t+1}^R \otimes \boldsymbol{\eta}_{t+1}^{S,X}$.*

Here, $\boldsymbol{\eta}_t^R$ is the predictive belief state of the agent T, and $\boldsymbol{\eta}_t^{S,X}$ is the belief state of the action-delayed agent U.

This product structure allows us to consider a simplified class of factored nonlinear models that produce the same statistics in a space with reduced dimensionality: Rather than representing the belief state as a product, we can preserve the contained information by expressing it as a direct sum $\boldsymbol{\eta}_t^{\text{fact}} \equiv \boldsymbol{\eta}_t^R \oplus \boldsymbol{\eta}_t^{S,X}$.

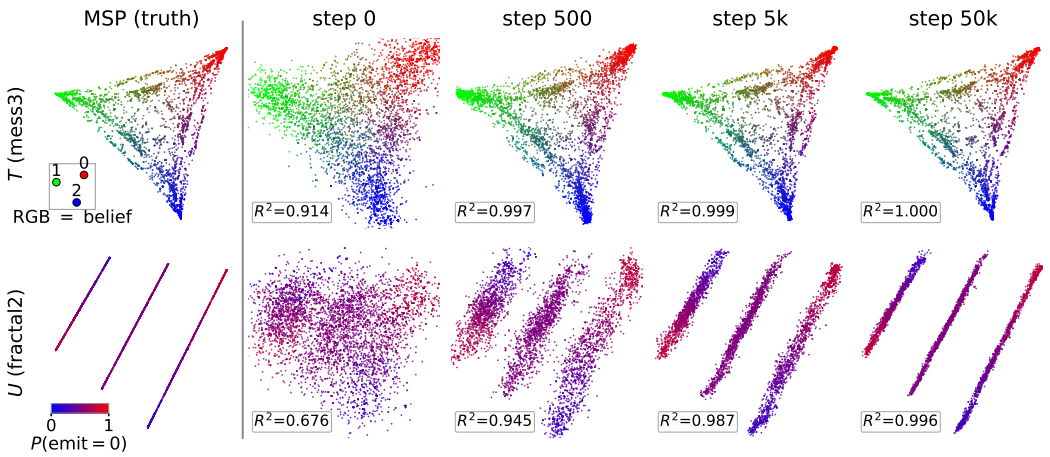


Figure 2: **In transformers trained on a coupled transducer process separable linear belief representations emerge over the first few hundred steps of training.** A coupled process with mess3 as agent T (top row) and fractal2 as agent U (bottom row). Leftmost column: Bayes-optimal belief MSP of each factor (top two PCA directions; colored by the belief resp. the emission probability). Remaining columns: linear reconstruction from the final-layer residual stream at training steps 0, 500, 5k and 50k (held-out linear probe refit per checkpoint, same PCA basis; held-out R^2 annotated). The role-swapped arrangement is shown in Figs. 9 and 10; protocol in Appendix K.

3.1 Factored Representation of Feedback Processes in Transformers

To validate the prediction of the Multi-Agent Belief Geometry Hypothesis, we demonstrate that a transformer trained on next-token prediction of the feedback process $p(x_{0:L}, y_{0:L})$ between two simple transducers represents the beliefs about the individual factors linearly in its residual stream. The transducers in this experiment are derived from the "mess3" (agent T) and "fractal2" families of HMMs (for details, see Appendix K). A linear probe fitting the residual stream activations after the last layer produces a faithful linear reconstruction of the Bayesian MSP of both agents T and U in the feedback loop which emerges over the first few hundred steps of training (see Fig. 2).

Note that while a linear probe on the randomly initialized model (step 0) already reaches R^2 of 0.83–0.92, the reconstructed geometry is visibly distorted; it improves over the first few hundred training steps to a near-perfect $R^2 \geq 0.996$ by step 50 000. The role-swapped setup with fractal2 taking precedence is presented in Appendix H.

3.2 Causal Steerability of Factored Belief Representations

So far, we have shown correlational evidence: as the model converges to the Bayes-optimal loss, linear representations of each agent's belief state are represented in the residual stream, generalizing what Shai et al. (2026) observed for conditionally independent factors arranged in a dependency chain to feedback loops. To test that these representations are also causally used by the transformer in next-token prediction, we intervened on the beliefs about either agent T or agent U separately by replacing their beliefs by beliefs from a counterfactual token sequence, and recorded whether the resulting predictions are close to the counterfactual Bayesian optimum (Fig. 3). We find that the KL divergence of the next-token prediction of the patched belief is slightly higher than, but close to, the KL divergence of the original setup, closing $\approx 98\%$ of the difference to a null comparison of the original prediction to the patched ground-truth or an unrelated control sequence (unprojected write; Appendix L compares an orthogonalized variant); the role-swapped loop closes the same fraction (Fig. 10).

Furthermore, the interventions show a dose-response relationship, with partial interventions producing intermediate predictions (see Fig. 17), indicating that the transformers learn the functional relationship between the belief and next token probabilities instead of memorizing. Note, however, that not all processes can be steered effectively (see Appendix L) with the reason for imperfect steerability of the "sns" process remaining an open subject for research.

4 MULTI-AGENT BELIEF DIMENSION

Having established that transformers encode beliefs about individual agents in coupled processes and that these factored representations are causally load-bearing, we next show that this encoding scheme comes with substantial embedding dimension advantages compared to an unfactored joint representation. We express this as the following hypotheses:

- **Multi-Agent Belief Dimension Hypothesis (MABDH):** A neural network that is trained on next-token prediction of a feedback process generated by the joint HMM V will converge to latent activation geometry that requires at most $D_{\text{fact}}(V)$ dimensions to express.
- **Coarse-Graining Hypothesis:** Coarse-graining the sequence can break this factorization advantage, leading to a higher required embedding dimension.

To test the dimensionality of transformer representations, we first define the product and factored dimensions.

Definition 1. *For an HMM T that can be factored into conditionally independent components*

$$\hat{T}^{(x)} \equiv \bigotimes_{n=0}^{N-1} \hat{T}(n)^{(x)}, \tag{7}$$

where $\hat{T}(n)$ operates on a vector space $\mathcal{R}(n)$:

- *The* product dimension $D_{prod}(T) = \prod_{n=0}^{N-1} |\mathcal{R}(n)| - 1$ *is the dimension of the simplex in which the linear mixed states live.*
- *The* factored dimension $D_{fact}(T) = \sum_{n=0}^{N-1} (|\mathcal{R}(n)| - 1)$ *is the dimension of the factored simplex.*

The nonlinear factored presentation of a process requires fewer dimensions to accurately predict the overall process.

Corollary 3. *For a joint HMM V composed of two interacting agents T with latent space $\mathcal{R}$ and U with latent space $\mathcal{S}$ and action space $\mathcal{X}$, the factored presentation of the feedback process uses a reduced dimension from the product presentation:*

$$D_{fact}(V) = \frac{|\mathcal{R}| + |\mathcal{S}||\mathcal{X}| - 2}{|\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1} D_{prod}(V). \tag{8}$$

Proof. Immediate from the definitions: the feedback process factors into T (space $\mathcal{R}$) and the action-delayed agent U (space $\mathcal{S} \otimes \mathcal{X}$), so $D_{\text{fact}}(V) = |\mathcal{R}| + |\mathcal{S}||\mathcal{X}| - 2$ while $D_{\text{prod}}(V) = |\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1$. □

Note that while any joint process can be described as a feedback loop between agents, Eq. (8) indicates when such a description is advantageous from an embedding dimension perspective. As the dimensionality advantage of the factored representation hinges on the dimensions of the factors, our claims should be interpreted about a process relative to a given compact decomposition: when the training data admit small factors, the network finds and exploits that structure.

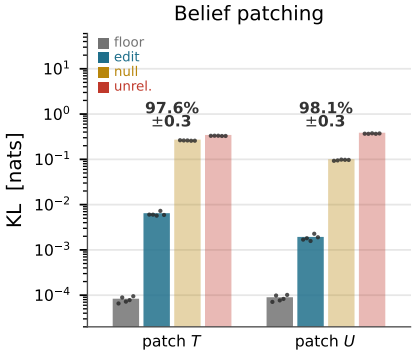


Figure 3: **Beliefs can be edited factor by factor.** Belief patching for the loop of Fig. 2: a counterfactual belief is spliced into one factor's residual-stream subspace, and the joint next-token distribution is compared in KL with the Bayes prediction for the patched beliefs (*edit*), the unedited model (*floor*), a write with no effect (*null*), and an unrelated sequence (*unrel.*). Annotated: fraction of the null-to-target gap closed (mean ± sd, five seeds). Protocol in Appendix L.

4.1 SOCIAL NETWORK

The dimensional reductions of belief states are particularly pronounced in a social network setting. In a social network, a collection of N subagents $\vec{A} \equiv \{A(0), A(1), \cdots A(N-1)\}$ interact with each other through a central social network hub H (see Fig. 4 a), mimicking a social media platform.

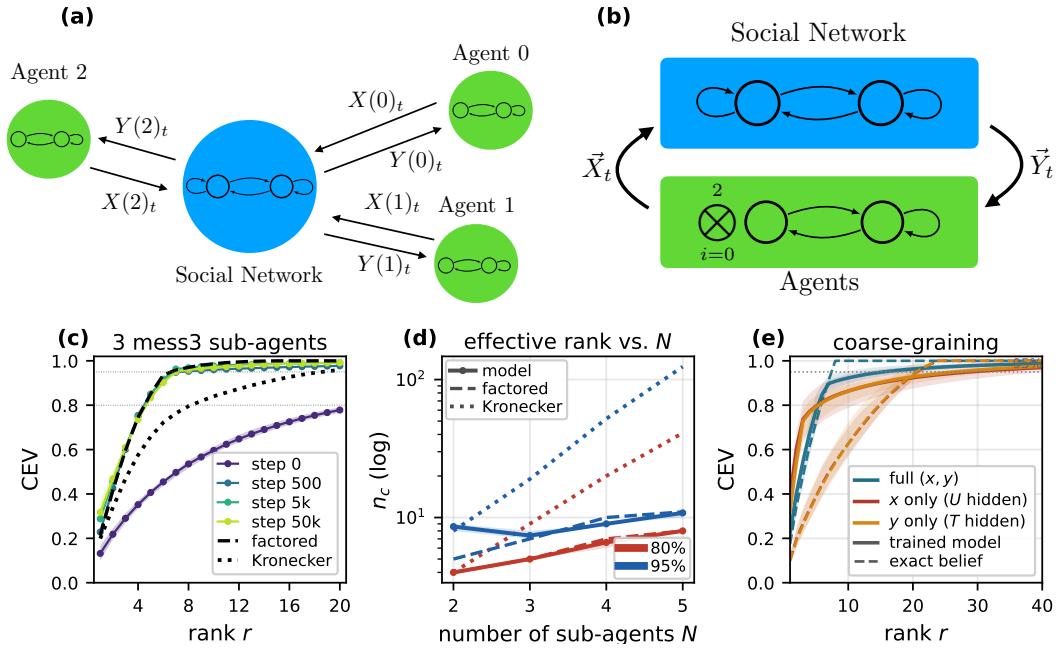


Figure 4: **The factored representation's dimensional advantage grows with the number of agents, and coarse-graining destroys it.** A social network in which a central hub automaton exchanges perceptions and actions with N agents **(a)** can be interpreted as a binary feedback loop in which the agents are compressed into a single macro-agent **(b)**. **(c)** Cumulative explained variance (CEV) of the final-layer residual stream follows factored prediction for the telephone process with three mess3 sub-agents. Solid curves: mean over five seeds (see Table 2); shaded bands: min–max across seeds. Dashed: CEV of the factored concatenation of the Bayes-optimal per-agent beliefs on the same sequences; dotted: CEV of their Kronecker product. **(d)** Effective rank n_c—the smallest number of principal components whose CEV reaches 80% (red) resp. 95% (blue)—for telephone processes with N mess3 sub-agents (BOS position excluded, 4096 sequences). Solid lines: the final-layer residual stream (mean over five seeds per N; the shaded min–max band never exceeds one component). Bayes-optimal beliefs themselves are the factored concatenation of the per-agent beliefs (dashed; simplex dimension $\sum_i (d_i - 1) = 8N$) and their Kronecker product (dotted; simplex dimension $\prod_i d_i - 1 = 9^N - 1$). For $N \geq 3$ the residual stream matches the factored beliefs to within one component at both thresholds, while the Kronecker beliefs require 2.5–$11\times$ more. **(e)** Coarse-graining forces a *higher*-dimensional representation: CEV by PCA rank, one color per observed interface, dashed lines the Bayesian-optimal belief and solid lines the trained transformer (mean over five independent draws of the process, one model per draw and interface; shaded bands: min–max across draws). Under full observation, the belief state factorizes, and the theory curve saturates at the 8-dimensional factored concatenation, whereas the coarse observers must track the entangled joint belief state (~ 22 dimensions). Process and model parameters in Appendix K.

Specifically, we suppose that each subagent $A(i)$ is imbued with memory $S(i)_t$, perceptions $Y(i)_t$, and actions $X(i)_t$, but that those actions only influence other agents through the central hub. The hub itself has memory R_t, which it uses to transform the vector of actions it receives from all subagents $\vec{X}_t \equiv \{X(0)_t, X(1)_t, \cdots X(N-1)_t\}$ to a perception vector $\vec{Y}_t \equiv \{Y(0)_t, Y(1)_t, \cdots Y(N-1)_t\}$ that it apportions to the subagents.

As derived in Appendix F, the factored representation of the social network feedback loop expresses the joint belief state as a direct sum of the social network belief (dimension $|\mathcal{R}| - 1$) and each action-delayed agent (dimension $|\mathcal{S}(i)||\mathcal{X}(i)| - 1$). Therefore, the factored presentation expresses the model in vastly reduced dimension:

$$D_{\text{fact}}(V) = (|\mathcal{R}| - 1) + \sum_{i=0}^{N-1} (|\mathcal{X}(i)||\mathcal{S}(i)| - 1), \tag{9}$$

which scales linearly with the number of agents N compared to the product dimension $D_{\text{prod}} = |\mathcal{R}| \prod_{i=0}^{N-1} |\mathcal{X}(i)||\mathcal{S}(i)| - 1$, which scales exponentially in N.

To investigate whether the predicted exponential dimensionality advantage is exploited by transformer networks, we consider a simplistic memoryless social network mimicking the children's game "telephone". In this setup, the social network simply passes the output of sub-agent $A(i)$ to sub-agent $A((i+1) \bmod N)$, meaning that $Y((i+1) \bmod N)_t = X(i)_t$. Each sub-agent is described by a transducer built from a mess3 process (see Fig. 14).

We trained a transformer on the joint process of the inputs and outputs of all agents $X(0:N)Y(0:N)$. The joint HMM has a dimension equal to d^N (with $d = 9$ for mess3), meaning that the mixed states live in a simplex of dimension $|\mathcal{X}|^N|\mathcal{S}|^N - 1 = d^N - 1$. However, because the joint machine has a product structure, the MSP can be projected into a space of dimension $N(|\mathcal{X}||\mathcal{S}| - 1) = N(d-1)$ or less. In our experiments, the CEV of the residual stream after the last layer follows the prediction by the factored hypothesis closely for the mess3 agents, and even exceeds it for the fractal2 process (Fig. 12). The reason for the lower dimensionality of representations in the fractal2 case is likely that the MSP of the fractal2 transducer only inhabits two one-dimensional sub-manifolds of the probability simplex, allowing for a more compressed representation.

For the mess3 agents, we further vary the number of agents between 2 and 5 and compare the residual stream's effective rank with that of the Bayes-optimal beliefs at the same CEV thresholds (Fig. 4d). For $N \geq 3$ the residual stream tracks the factored beliefs to within one component, whereas the Kronecker beliefs need 2.5–11$\times$ more components at 95% (19, 52 and 124 for $N = 3, 4, 5$).

4.2 Complexification Through Coarse-Graining

Surprisingly, a process can become more complex, in this case meaning that it requires more dimensions to be modeled predictively, when you coarse-grain it by omitting either agent's actions from the token stream. When coarse-graining agent U's actions, the resulting linear operator for the Y process is $\hat{W}^{(y)} \equiv \sum_x \hat{V}^{(xy)} = \sum_x \hat{T}^{(y|x)} \otimes \hat{\mathcal{U}}^{(x|y)}$. Here, summing over x entangles the operators so that they are no longer necessarily separable into a tensor product. The output marginalized operator now lives in the $(|\mathcal{S}||\mathcal{R}||\mathcal{X}| - 1)$-simplex, potentially increasing the size of the latent space necessary to predict that process beyond the dimension of the factored presentation.

We provide evidence for this complexification in transformers predicting a coarse-grained process. In this setup, the state transitions of each agent are given by an epsilon-noisy, input-conditional permutation of its previous states, with tokens generated noisily depending on only the arriving state with a token alphabet that is strictly smaller than the number of internal states (for details, see Appendix M). Exact filtering shows that the fully observed process admits an 8-dimensional factored belief state, whereas either coarse-graining forces an entangled joint belief state of $\sim$22 dimensions. The trained transformers follow suit: the final-layer residual stream of the fully observing models crosses the 95% CEV threshold at rank 16 on average (five independent draws of the process, range 10–21), while the coarse-grained models cross it only at rank 27 on average (range 12–42), and in every one of the ten within-draw comparisons the coarse model's rank exceeds that of its fully observing counterpart (see Fig. 4e and Appendix M.2).

5 Discussion

In this manuscript, we introduce a theory of prediction of coupled processes, explaining both the theoretical advantages and empirical evidence for transformers representing belief states about coupled processes in a factorized manner. In particular, we prove that any joint process can be viewed as a feedback process between multiple agents, and that this allows the beliefs about the states of the agents to factorize, which can lead to an exponential dimensional advantage in their embedding dimension. We further demonstrate how, counterintuitively, coarse-graining a process by hiding the emissions of one of these subagents can lead to a higher required dimensionality of the belief space, a "complexification" despite less information to predict. In experiments, we show how transformers are able to infer when tokens from a sequence originate from a coupled process, and express their lower-dimensional factored representations. We furthermore show how these belief representations can be individually steered by patching counterfactual beliefs.

Related work. Our work sits at the intersection of different lines of work. *First*, the formal study of belief-state geometry: RNNs and transformers trained on next-token prediction linearly represent Bayesian beliefs (Shai et al., 2024; Riechers et al., 2025; Piotrowski et al., 2025), and

beliefs about independent or conditionally independent processes factorize, yielding an exponential dimensional advantage (Shai et al., 2026). This line of work develops the idea that one can describe the computation of transformers at inference time using belief automata (Shai et al., 2024), cf. (Zhang et al., 2024; Adriaensen & Maene, 2024) for other kinds of automata, which in the case of agents' belief states can be interpreted in light of established belief MDPs approaches (Åström, 1965; Kaelbling et al., 1998; Subramanian et al., 2022; Biehl & Virgo, 2023; Rosas et al., 2025; Baltieri & Virgo, 2026). We extend this to the general case of joint processes read as feedback loops between coupled agents.

Second, emergent world models: Decodable world models (Li et al., 2023) can be incoherent (Vafa et al., 2024) or not maintained (Pereira & Zuidema, 2026), which is the gap that our selectivity controls and causal interventions are designed to close (Hewitt & Liang, 2019; Elazar et al., 2021), following the patching methodology of Zhang & Nanda (2024).

Third, factored representations: Unsupervised disentanglement is impossible without inductive bias (Locatello et al., 2019). Our framework suggests such a bias towards a derived target geometry. This is similar, in spirit, to the interaction-based disentanglement approach of Thomas et al. (2018). In the context of computational mechanics applications to reinforcement learning, this factorization is discussed (Baltieri et al., 2026) with connections to predictive state representations (Littman et al., 2001; Singh et al., 2004) (rather than belief MDPs) and their counterparts. More generally, factored representations in RL have been used to encode structure that can be used to create more compact and modular world models (Boutilier & Poole, 1996; Singh et al., 2021), but their possible applications in the context of interpretability of transformer models remain underexplored.

Fourth, multi-agent systems and beliefs: Andreas (2022) and Zhu et al. (2024) argue that language models infer and linearly encode latent interacting agents. Here we make the corresponding geometry and dimension explicit. An extended discussion is in Appendix C.

Limitations and future work. While our work shows the theoretical advantages and transformers' preference for factored representations, our empirical evidence is as of yet limited to small transformers trained directly on synthetic feedback processes with known ground truths. Devising methods to discover beliefs in networks where the ground truth beliefs are withheld—sparse dictionary learning being a natural candidate (Huben et al., 2024; Templeton et al., 2024)—is the pivotal next step towards determining whether pretrained language models likewise take advantage of factored belief representations. Furthermore, since a given process may admit multiple factorized representations, which one of these a transformer intrinsically favors remains an open question.

Note also that the Multi-Agent Belief Dimension Hypothesis only provides an upper bound to the encoding dimensionality of the beliefs. For processes whose belief states fill the simplex, like the mess3 process, we find this bound to be tightly met by the PCA dimension of trained transformers, while for processes with more sparse belief states like fractal2, the transformers tend to find even more parsimonious representations. Analyzing this minimal intrinsic belief dimension of processes to find a tighter bound is an open question. Note also that variance-based rank measures cannot count features stored in superposition (Elhage et al., 2022); the reported PCA ranks therefore bound the linearly readable belief code, and the dimensional evidence is corroborated by the geometric and causal analyses.

The framework described in this paper can be extended in multiple ways. First, the simple feedback or telephone architecture of the coupled processes can be generalized to more complex interaction graphs between agents, raising the question whether a transformer network would still be able to infer this graph structure from sequence data.

Second, a natural extension of this work would involve an investigation of the causal power of the factored, or disentangled representations reconstructed from the transformer residual stream, in the tradition that portrays factorization as a necessary component of causal reasoning (Schölkopf et al., 2021) and causal learning (Torresan & Baltieri, 2024).

Overall, by establishing a theory of prediction for coupled processes rooted in first principles of computational mechanics, this work lays the foundation to investigate whether transformers' next token prediction can be described as arising from feedback loops between multiple subcomponents—composite minds of interacting agents. This modular view promises to provide a powerful harness for AI safety, allowing us to subdivide monolithic high-dimensional models into subcomponents which can be interpreted, audited and steered one at a time.

AI USE STATEMENT

In this work, we used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, idea generation, for writing and debugging code, reviewing mathematical proofs and for drafting and editing text.

We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, assistance with translation, supporting qualitative and thematic data analysis or result interpretation.

Generating synthetic datasets or cleaning and reformatting datasets are not applicable to this work.

All AI-assisted work was reviewed by the authors: generated code was verified and tested, and generated text was edited for accuracy. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We submitted the code to reproduce our findings as supplementary material and will make it publicly accessible for the final release.

REFERENCES

David Abel. A theory of abstraction in reinforcement learning. *arXiv preprint arXiv:2203.00397*, 2022.

Rik Adriaensen and Jaron Maene. Extracting finite state machines from transformers. *arXiv preprint arXiv:2410.06045v1*, 2024. URL https://arxiv.org/abs/2410.06045v1.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, 2022.

W. Ross Ashby. *An Introduction to Cybernetics*. Chapman and Hall, London, 1956.

W. Ross Ashby. *Design for a Brain: The Origin of Adaptive Behavior*. Chapman and Hall, London, second edition, 1960.

Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965.

Alessandro Attanasi, Andrea Cavagna, Lorenzo Del Castello, Irene Giardina, Tomas S Grigera, Asja Jelić, Stefania Melillo, Leonardo Parisi, Oliver Pohl, Edward Shen, et al. Information transfer and behavioural inertia in starling flocks. *Nature physics*, 10(9):691–696, 2014.

Manuel Baltieri and Nathaniel Virgo. Bayesian updates from coalgebraic determinisation. *arXiv preprint arXiv:2607.00034*, 2026. URL https://arxiv.org/abs/2607.00034.

Manuel Baltieri, Martin Biehl, Matteo Capucci, and Nathaniel Virgo. A Bayesian interpretation of the internal model principle. *arXiv preprint arXiv:2503.00511*, 2025.

Manuel Baltieri, Filippo Torresan, Yivan Zhang, Alexander Boyd, and Fernando E. Rosas. World models of environment, agent and joint agent-environment systems. *arXiv preprint arXiv:2608.20401*, 2026. URL https://arxiv.org/abs/2608.20401.

N. Barnett and J. P. Crutchfield. Computational mechanics of input-output processes: Structured transformations and the ϵ-transducer. *J. Stat. Phys.*, 161(2):404–451, 2015. doi: 10.1007/s10955-015-1327-5. URL http://dx.doi.org/10.1007/s10955-015-1327-5. SFI Working Paper 14-12-046; arxiv.org: 1412.2690 [cond-mat.stat-mech].

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Martin Biehl and Nathaniel Virgo. Interpreting systems as solving POMDPs: A step towards a formal understanding of agency. In *Active Inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, Revised Selected Papers*, pp. 16–31. Springer, 2023. doi: 10.1007/978-3-031-28719-0_2.

Craig Boutilier and David Poole. Computing optimal policies for partially observable decision processes using compact representations. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp. 1168–1175. AAAI Press, 1996.

A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18:023049, 2016. doi: doi:10.1088/1367-2630/18/2/023049. SFI Working Paper 15-07-025; arxiv.org:1507.01537 [cond-mat.stat-mech].

A. B. Boyd, D. Mandal, and J. P. Crutchfield. Thermodynamics of modularity: Structural costs beyond the Landauer bound. *Phys. Rev. X*, 8(031036), August 2018.

Alexander Boyd, Franz Nowak, David Hyland, Manuel Baltieri, and Fernando E Rosas. From monoliths to modules: Decomposing transducers for efficient world modelling. *arXiv preprint arXiv:2512.02193*, 2025.

Sasha Brenner, Thomas R. Knösche, and Nico Scherf. Predictive statistics shape emergent world representations of grid walkers. *arXiv preprint arXiv:2603.16689*, 2026.

Jelle Bruineberg, Krzysztof Dołęga, Joe Dewhurst, and Manuel Baltieri. The emperor's new Markov blankets. *Behavioral and Brain Sciences*, 45:e183, 2022. doi: 10.1017/S0140525X21002351.

C. J. Burke and M. Rosenblatt. A Markovian function of a Markov chain. *The Annals of Mathematical Statistics*, 29(4):1112–1122, 1958. doi: 10.1214/aoms/1177706444.

Andrea Cavagna, Alessio Cimarelli, Irene Giardina, Giorgio Parisi, Raffaele Santagati, Fabio Stefanini, and Massimiliano Viale. Scale-free correlations in starling flocks. *Proceedings of the National Academy of Sciences*, 107(26):11865–11870, 2010.

Bilian Chen, Biyang Ma, Yifeng Zeng, Langcai Cao, and Jing Tang. Tensor decomposition for multi-agent predictive state representation. *arXiv preprint arXiv:2005.13706*, 2020.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012. doi: 10.1038/NPHYS2190.

J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989. doi: 10.1103/PhysRevLett.63.105.

J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016. Santa Fe Institute Working Paper 13-09-028; arXiv:1309.3792 [cond-mat.stat-mech].

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? Inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*, 2021.

Tabea Dreyer, Amir Haluts, Amos Korman, Nir Gov, Ehud Fonio, and Ofer Feinerman. Comparing cooperative geometric puzzle solving in ants versus humans. *Proceedings of the National Academy of Sciences*, 122(1):e2414274121, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Lukas J Fiderer, Paul C Barth, Isaac D Smith, and Hans J Briegel. The work capacity of channels with memory: Maximum extractable work in percept-action loops. *arXiv preprint arXiv:2504.06209*, 2025.

Daniel Filan, Stephen Casper, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell. Clusterability in neural networks. *arXiv preprint arXiv:2103.03386*, 2021.

Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 122–130, 2018.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

K. Friston. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.*, 11:127–138, 2010.

Karl Friston. Life as we know it. *Journal of The Royal Society Interface*, 10(86):20130475, 2013. doi: 10.1098/rsif.2013.0475.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187, 2024.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003. doi: 10.1016/S0004-3702(02)00376-4.

Piotr J. Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005. doi: 10.1613/jair.1579.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. Brief independent investigation of agents' behavior, reasoning and collaboration in the openai / hugging face hacking incident. Report, Model Evaluation and Threat Research (METR), 2026. URL https://metr.org/hugging-face-incident-report-aug-2026.pdf.

M. Gu, K. Wiesner, E. Rieper, and V. Vedral. Quantum mechanics can reduce the complexity of classical models. *Nature Comm.*, 3(762):1–5, 2012. doi: 10.1038/ncomms1761.

Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A. Pires, Tobias Pohlen, and Rémi Munos. Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*, 2018.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2733–2743, 2019.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

Edward S. Hu, Kwangjun Ahn, Qinghua Liu, Haoran Xu, Manan Tomar, Ada Langford, Dinesh Jayaraman, Alex Lamb, and John Langford. The belief state transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

David L Hull. *Science as a process: An evolutionary account of the social and conceptual development of science*. University of Chicago Press, 1988.

R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011. doi: 10.1063/1.3637494. URL http://arxiv.org/abs/1105.2988.

Alexandra M Jurgens and James P Crutchfield. Ambiguity rate of hidden markov processes. *Physical Review E*, 104(6):064107, 2021a.

Alexandra M Jurgens and James P Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden markov processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(8):083114, 2021b.

Alexandra M Jurgens and James P Crutchfield. Shannon entropy rate of hidden markov processes. *Journal of Statistical Physics*, 183(2):32, 2021c.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All else being equal be empowered. In *Advances in Artificial Life: 8th European Conference (ECAL)*, volume 3630 of *Lecture Notes in Computer Science*, pp. 744–753, 2005.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

David Krakauer, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nihat Ay. The information theory of individuality. *Theory in Biosciences*, 139(2):209–223, 2020.

Michael A. Lepori, Thomas Serre, and Ellie Pavlick. Break it down: Evidence for structural compositionality in neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42623–42660, 2023.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023.

Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, New York, third edition, 2008. doi: 10.1007/978-0-387-49820-1.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

J. L. Massey. Causality, feedback and directed information. In *Proc. 1990 Intl. Symp. on Info. Th. and its Applications, Waikiki, Hawaii, Nov. 27-30*, pp. 1–6, 1990.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Marvin Minsky. *The Society of Mind*. Simon and Schuster, 1986.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Tianwei Ni, Benjamin Eysenbach, Erfan Seyedsalehi, Michel Ma, Clement Gehring, Aditya Mahajan, and Pierre-Luc Bacon. Bridging state and history representations: Understanding self-predictive RL. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/666c1861d709bd84e20b6e0e02a2c223-Abstract-Conference.html.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. Transformer Circuits Thread, 2022. URL https://www.transformer-circuits.pub/2022/mech-interp-essay.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Devin Pereira and Willem Zuidema. Transformers struggle to use their emergent world models: Revisiting the tower of hanoi, and the illusion of thinking. *arXiv preprint arXiv:2608.07077*, 2026.

Mateusz Piotrowski, Paul M Riechers, Daniel Filan, and Adam S Shai. Constrained belief updates explain geometric structures in transformer representations. *arXiv preprint arXiv:2502.01954*, 2025.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-markov decision processes. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pp. 1011–1016, 2003.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/richens25a.html.

Paul M Riechers, Thomas J Elliott, and Adam S Shai. Neural networks leverage nominally quantum and post-quantum representations. *arXiv preprint arXiv:2507.07432*, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Fernando Rosas, Alexander Boyd, and Manuel Baltieri. AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability. *arXiv preprint arXiv:2504.04608*, 2025.

Fernando E. Rosas, Pedro A. M. Mediano, Henrik J. Jensen, Anil K. Seth, Adam B. Barrett, Robin L. Carhart-Harris, and Daniel Bor. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLoS Computational Biology*, 16(12):e1008289, 2020.

J. Ruebeck, R. G. James, John R. Mahoney, and J. P. Crutchfield. Prediction and generation of binary Markov processes: Can a finite-state fox catch a markov mouse? *CHAOS*, 28:013109, 2018. doi: https://doi.org/10.1063/1.5003041. SFI Working Paper 2017-08-027; arxiv.org:1708.00113 [cond-mat.stat-mech].

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? On the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780, 2022.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Adam Shai, Loren Amdahl-Culleton, Casper L Christensen, Henry R Bigelow, Fernando E Rosas, Alexander B Boyd, Eric A Alt, Kyle J Ray, and Paul M Riechers. Transformers learn factored representations. *arXiv preprint arXiv:2602.02385*, 2026.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander G Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. *Advances in Neural Information Processing Systems*, 37:75012–75034, 2024.

C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Gautam Singh, Skand Vishwanath Peri, Junghyun Kim, Hyunseok Kim, and Sungjin Ahn. Structured world belief for reinforcement learning in POMDP. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9744–9755. PMLR, 2021.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 512–519, 2004.

Susanne Still. Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85 (2):28005, 2009.

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022. URL https://jmlr.org/papers/v23/20-1165.html.

Sekhar Tatikonda and Sanjoy Mitter. The capacity of channels with feedback. *IEEE Transactions on Information Theory*, 55(1):323–349, 2009. doi: 10.1109/TIT.2008.2008147.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Jayden Teoh, Manan Tomar, Kwangjun Ahn, Edward S. Hu, Tim Pearce, Pratyusha Sharma, Akshay Krishnamurthy, Riashat Islam, Alex Lamb, and John Langford. Next-latent prediction transformers learn compact world models. *arXiv preprint arXiv:2511.05963*, 2025.

Valentin Thomas, Emmanuel Bengio, William Fedus, Jules Pondard, Philippe Beaudoin, Hugo Larochelle, Joelle Pineau, Doina Precup, and Yoshua Bengio. Disentangling the independently controllable factors of variation by interacting with the world. *arXiv preprint arXiv:1802.09484*, 2018. URL https://arxiv.org/abs/1802.09484.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-Action Cycle: Models, Architectures, and Hardware*, pp. 601–636. Springer, 2011. doi: 10.1007/978-1-4419-1452-1_19.

Filippo Torresan and Manuel Baltieri. Disentangled representations for causal cognition. *Physics of Life Reviews*, 51:343–381, 2024. doi: 10.1016/j.plrev.2024.10.003.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Daniel Ray Upper. *Theory and algorithms for hidden Markov models and generalized hidden Markov models*. University of California, Berkeley, 1997.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Nathaniel Virgo, Martin Biehl, Manuel Baltieri, and Matteo Capucci. A "good regulator theorem" for embodied agents. In *Artificial Life Conference Proceedings 37 (ALIFE 2025)*, pp. 46. MIT Press, 2025. arXiv:2508.06326.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Amy Zhang, Zachary C. Lipton, Luis Pineda, Kamyar Azizzadenesheli, Anima Anandkumar, Laurent Itti, Joelle Pineau, and Tommaso Furlanello. Learning causal state representations of partially observable environments. *arXiv preprint arXiv:1906.10437*, 2019. URL https://arxiv.org/abs/1906.10437.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Yihao Zhang, Zeming Wei, and Meng Sun. Automata extraction from transformers. *arXiv preprint arXiv:2406.05564*, 2024. URL https://arxiv.org/abs/2406.05564.

Yivan Zhang, Ziyan Luo, and Manuel Baltieri. Compositional behavioral semantics for state abstraction in reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.25357. Also available as arXiv:2606.25357.

Zhi Zhang, Zhuoran Yang, Han Liu, Pratap Tokekar, and Furong Huang. Reinforcement learning under a multi-agent predictive state representation model: Method and theory. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PLDOnFoVm4.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*, 2024.

Kevin JS Zollman. The communication structure of epistemic communities. *Philosophy of science*, 74(5):574–587, 2007.

A Computational Mechanics Background

A.1 Interfaces: Agent Behavior

Collective intelligence arises from the interaction of individual *agents*. We define an agent's "behavior" as what is observable from the outside: the agent's actions in response to its external stimulus. In other words, what sequence of actions $y_{0:L} \equiv y_0 y_1 \cdots y_{L-1}$ will that agent take in response to the sequence of perceptions $x_{0:L} = x_0 x_1 \cdots x_{L-1}$?

Formally, we characterize behavior as the stochastic mapping from the agent's sequence of perceptions $x_{0:L}$ to its actions $y_{0:L}$, otherwise known as an *interface*. Specifically, an interface is a set of conditional probabilities that maps ever-longer perception sequences to action sequences of the same length:

$$I[Y|X] \equiv \{p(Y_{0:L}|X_{0:L}) \\ |L \in \{0, 1, 2 \cdots\}\},$$

where X_t and Y_t are the random variables at time t for the perception and action, respectively, while x_t and y_t are specific realizations. The probability $I[y_{0:L}|x_{0:L}] \equiv p(Y_{0:L} = y_{0:L}|X_{0:L} = x_{0:L})$ reflects the probability of performing the action sequence $y_{0:L}$ if the agent receives the specific perception sequence $x_{0:L}$ in a feed-forward context (Barnett & Crutchfield, 2015; Boyd et al., 2025). This interface construction is sufficiently general to encode *any* stochastic strategy.

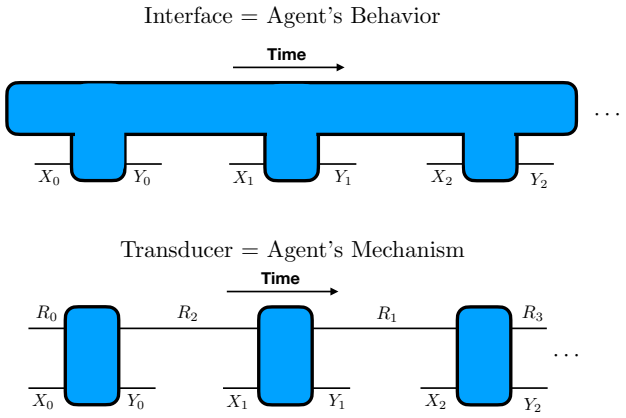


Figure 5: (TOP) An agent's behavior is described by the interface it generates, which can be thought of as a channel whose inputs and outputs are arranged in time. (BOTTOM) An agent's mechanism is described by a transducer, which describes a recursive circuit that uses a memory random variable R_t to carry information between past and future.

We note one constraint on these interfaces to make them realizable: they must be causal. This means future perceptions cannot affect past actions (Boyd et al., 2025), meaning that

$$p(Y_{0:L} = y_{0:L}|X_{0:\infty} = x_{0:\infty}) = p(Y_{0:L} = y_{0:L}|X_{0:L} = x_{0:L}). \tag{10}$$

This is known as being non-anticipatory or anticipation-free (Boyd et al., 2025; Barnett & Crutchfield, 2015). In essence, the strategy an agent employs must be a function of the past. If it could use its future perceptions to choose its actions, this would be like writing down tomorrow's lottery numbers today to obtain a vast fortune: it violates our understanding of time's arrow. Fig. 5 shows how the agent's behavior can be conceived of as a circuit in time. The ordering of elements prevents the input future $\overrightarrow{X}_t \equiv X_t X_{t+1} \cdots$ from influencing the output past $\overleftarrow{Y}_t \equiv Y_0 \cdots Y_{t-1}$.

While the form of the channel remains fixed in a feedback context, the true conditional probability one observes between X and Y can change. In the base case, where another agent (agent U) takes the action sequence $y_{0:L}$ of our original agent (agent T) as its own perceptions, then returns its actions $x_{1:L+1}$ to agent T, the joint sequence probability of these agents acting in sequence is given by the product of their interfaces (Fiderer et al., 2025):

$$\Pr(X_{0:L} = x_{0:L}, Y_{0:L} = y_{0:L}) = I_T[y_{0:L}|x_{0:L}] I_U[x_{0:L}|y_{0:L}]. \tag{11}$$

Consider two very simple agents. Agent T produces a 0 or 1 as its action with equal probability at all times, and Agent U is simply the identity channel, meaning that $x_{t+1} = y_t$. In this case $\Pr(Y_{0:L} = y_{0:L}|X_{0:\infty} = x_{0:\infty}) = \delta_{y_{0:L}, x_{1:L+1}}$, meaning that the conditional probability from perceptions X to actions Y is *anticipatory*. It should be noted that this is not because a feedback loop somehow violates causality or time's arrow. What looks like anticipation of future perceptions by present actions is simply those actions becoming the future perceptions. Therefore, the true conditional probability of the feedback process is not necessarily the same as the conditional probability that defines the feedback loop

$$\Pr(Y_{0:L} = y_{0:L}|X_{0:L} = x_{0:L}) \neq I_T[y_{0:L}|x_{0:L}]. \tag{12}$$

Equality only holds for feed-forward networks of transducers.

In summary, to examine the effect an agent has on other agents in its environment, one needs only determine its interface $I[Y|X]$. While it won't be equivalent to the conditional probability in the case of feedback, it can be used to directly calculate the feedback process when it is combined with the interface of its feedback companion. An agent's causal interface is its objective statistical behavior.

A.2 Transducers: Agent Mechanism

Alongside its perception X_t and action Y_t at time t, an agent also has an internal (latent) "state of mind" that reflects how it thinks and responds to its environment. Let us represent this latent random variable as R_t. Fig. 5 shows how the dynamics of the latent, together with the perceptions and actions, specify the mechanism of this agent. The present perception X_t and latent R_t specify both the current action Y_t and next latent R_{t+1}. For instance, if someone thinks "it is nighttime" and perceives a ray of sunshine on the horizon, their next thought will likely be along the lines "it is dawn," and perhaps they will act by making coffee.

The most general way to capture this mechanism is to specify a conditional probability

$$T^{(y|x)}_{r\to r'} = \Pr(Y_t = y, R_{t+1} = r' | X_t = x, R_t = r) \tag{13}$$

of the next state r' and action y given the present state r and perception x. This specifies a *transducer* automaton (Barnett & Crutchfield, 2015; Boyd et al., 2025; 2016) that operates on the latent space $\mathcal{R}$ with a symbol-labeled linear operator[2]

$$\hat{T}^{(y|x)} \equiv \sum_{r,r'} T^{(y|x)}_{r\to r'} |r'\rangle\langle r|. \tag{14}$$

Partially Observable Markov Decision Processes (POMDPs), a standard model for agentic dynamics, are a type of transducer. However, these POMDPs obey restrictions that limit the dependencies between perception, mind, and action (Rosas et al., 2025). By contrast, transducer models are fully general, because every transducer produces a causal interface through the relation

$$I_T[y_{0:L}|x_{0:L}] = \langle 1| \prod_{t=0}^{L-1} \hat{T}^{(y_t|x_t)} |\rho_0\rangle_R, \tag{15}$$

and for every causal interface, there exists a transducer that produces that channel. Here, the bra and ket on either side of the expression represent row and column vectors in the space of $\mathcal{R}$

$$\langle 1| \equiv \sum_r \langle r| \tag{16}$$

$$|\rho_0\rangle_R \equiv \sum_r |r\rangle \Pr(R_0 = r). \tag{17}$$

In other words, the class of transducers fully expresses all possible agentic behaviors.

For every causal behavior, there exists a transducer model that produces that behavior and vice versa, but that doesn't mean the mapping is one-to-one. In fact, for every interface, there are infinitely many transducer mechanisms which yield that particular perception-to-action mapping. Each such transducer is known as a *presentation* of an interface.

The multiplicity of transducer presentations is a reflection of the fact that there are many mechanisms to generate the same observable statistics. Consider, for example, a help window on a website. The user experience of the statistical behavior of that interface might be the same regardless of whether it's a person on the other side, a lookup table for responses, or an AI.But the latent "mind" of the agent driving those responses would be very different. The presentation captures dynamics of the latent in parallel to the perceptions and actions: the mechanism. These transducer presentations parallel the role of Hidden Markov Models (HMMs) in modeling a stochastic process (Upper, 1997; Crutchfield & Feldman, 2003; Shalizi & Crutchfield, 2001).

[2]We use the bra-ket notation of quantum mechanics, where $|r\rangle$ represents a column vector with a 1 in the rth place, and zeros everywhere else.

A.3 Mixed State Presentation: Universal Agent Beliefs

The parallel between HMMs and transducers extends further to their predictive belief states. Among the many presentations of a process or interface, predictive presentations, which use the past to anticipate future outcomes, have garnered special attention for both their theoretical implications in computational mechanics (James et al., 2011; Boyd et al., 2018; Crutchfield et al., 2016; Crutchfield, 2012), and their application to LLMs at the frontier of AI interpretability (Shai et al., 2024; 2026; Piotrowski et al., 2025). Specifically, it is possible to compute the predictive states of a belief-transducer by examining the Mixed State Presentation (MSP) (Jurgens & Crutchfield, 2021c;b; Rosas et al., 2025), also known as belief states (Shai et al., 2026; 2024):

$$|\rho_{t+1}\rangle_R = \frac{\hat{T}^{(y_t|x_t)}|\rho_t\rangle_R}{\langle 1|\hat{T}^{(y_t|x_t)}|\rho_t\rangle_R}. \tag{18}$$

This belief state provides the distribution of latent states inferred from the action|perception history $y_{0:t}|x_{0:t}$:

$$|\rho_t\rangle_R \equiv \sum_r |r\rangle \Pr(R_t = r|Y_{0:t} = y_{0:t}, X_{0:t} = x_{0:t}).$$

The predictive belief states linearly encode the information from the past that is necessary to predict the token. Unless there is redundancy in the presentation, the MSP is isomorphic to the ϵ-machine or ϵ-transducer: unique minimal prediction presentations of the process (Barnett & Crutchfield, 2015; Crutchfield & Feldman, 2003). These objects provide predictions for the geometry of activations within neural networks trained on next-token prediction (Shai et al., 2024; Piotrowski et al., 2025). In this way, the belief geometry reflects a sort of universal Platonic representation of the agent's state of mind. In what follows, we will extend these results to many agents exchanging information with each other, showing that their individual belief-states are components of the belief-states of the whole.

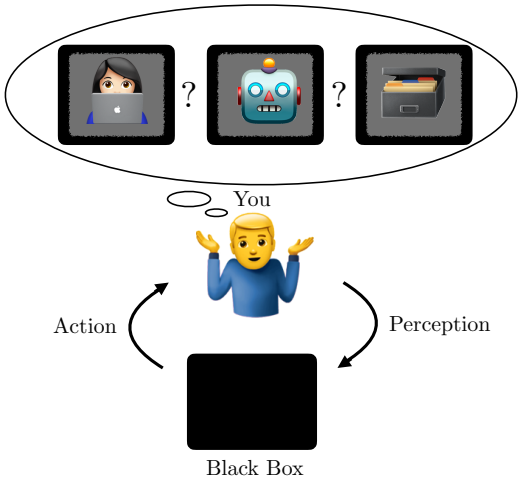


Figure 6: An agent's identity may be obscured by a black box. Even if you can perfectly characterize the way they map their perceptions to actions, you may be uncertain as to how their latent structure leads to those actions. This is a classic manifestation of the Turing test, where many mechanisms (human, AI, lookup tables) can all produce statistically similar behavior.

B Beyond Feed-forward

Unlike feed-forward networks of transducers (Boyd et al., 2025; Shai et al., 2026), rather than only using the memory space of the component processes, this construction also includes the perception space of Agent T as memory. Some additional memory is necessary to carry additional information in the feedback loop forward in time.

Consider the simple example of two memoryless agents, which simply stochastically transform their perceptions to actions

$$T^{(y|x)} = \Pr(Y_t = y|X_t = x) \tag{19}$$

$$U^{(x|y)} = \Pr(X_{t+1} = x|Y_t = y). \tag{20}$$

In the feed-forward case, in which we break the feedback $U^{(x|y)} = U^{(x)}$, we simply obtain an IID process, which is also memoryless. No information is carried between time steps, and the joint process can be expressed as an uncorrelated product

$$\Pr(X_{0:L}, Y_{0:L}) = \prod_{t=0}^{L-1} U^{(x_t)} T^{(y_t|x_t)}. \tag{21}$$

By contrast, if we allow for feedback, then we allow for the Markovian dependency. In fact, in the marginalized sub-processes $\Pr(X_{0:L})$ or $\Pr(Y_{0:L})$, *any* Markov order-1 process can be realized by

simply making Y_t a copy of X_t. Therefore, we are able to introduce memory, temporal correlations, and complexity into a collection of maximally simple memoryless agents by linking them into a feedback process. In essence, *complexity emerges* through feedback. This is qualitatively different than feed-forward processes, where the complexity of the individual components cannot exceed the sum of the parts. The following sections explore the complexity of feedback processes in greater detail through the dimension of their mixed states.

Feedback processes clearly go beyond the feedforward processes, but how much further do they go? Are there some systems that can be naturally understood as separate agents interacting while it is impossible to fully distinguish separate agents in other scenarios? For instance, it may be fairly clear that we can distinguish each of the people in a country as a distinct agent with their own beliefs, while the country itself is a macromind that operates through the composition of the actions and perceptions of all its members. On the other hand, can we model every individual cell in our body as an individual agent?

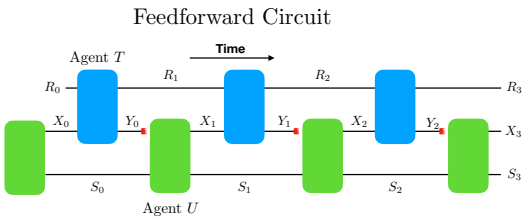


Figure 7: A feedforward circuit composed of multiple agents (bottom) is a special case of a feedback loop. In this case, that is achieved by eliminating the dependence of Agent U on Agent T's action, indicated by striking the wire for Y_t going to U with a red slash.

Feedforward processes provide a firm binary, identifying which sort of probabilistic dependencies allow for an interpretation as a feedforward network of agents. This boils down to whether each agent is anticipation-free from its actions to its perceptions (Barnett & Crutchfield, 2015). However, no such binary appears to exist in a network of agents linked through feedback. This is because, as shown in Appendix G, *any* joint process can be decomposed into a feedback process. Therefore, the results described in what follows are relevant to *any* joint process. There will always be a way to interpret the system as if it were two or more agents interacting. Whether such an interpretation is *useful*, however, is a quantitative question which depends on the complexity of the individual factors.

C Extended Related Work

Predictive states and belief representations. The object we track—the minimal sufficient statistic of the past for predicting the future—has been arrived at independently several times. Computational mechanics identifies it with the causal states of the ϵ-machine (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001; Crutchfield, 2012), extended to input-driven systems as ϵ-transducers (Barnett & Crutchfield, 2015) and to generalized HMMs by Upper (1997). Reinforcement learning arrived at the same place via belief-MDPs (Åström, 1965; Kaelbling et al., 1998) and predictive state representations (Littman et al., 2001; Singh et al., 2004), which define state by what it predicts rather than by a postulated latent. Learned analogs appear in neural predictive belief representations (Guo et al., 2018) and in world-model agents (Ha & Schmidhuber, 2018; Hafner et al., 2023). We take the computational mechanics perspective here, in which the belief simplex, its dimension, and the update operator are all available in closed form, making representational claims quantifiable. Rosas et al. (2025) use this framing to derive fundamental limits on world models for agent sandboxing and interpretability.

Belief geometry in trained networks. Shai et al. (2024) first showed that the residual stream of a small transformer trained on next-token prediction contains a linear image of the process's mixed-state presentation, including its fractal structure; Piotrowski et al. (2025) and Riechers et al. (2025) sharpen when and where this occurs. Shai et al. (2026) propose and test the factored world hypothesis: for a process built from independent sub-processes, or from a dependency chain of sub-processes that remain conditionally independent given the emitted tokens, the network represents the direct sum of per-factor beliefs rather than the tensor product, an exponential saving in embedding dimension. Our contribution is to extend their framework to the more general case of feedback processes. Concurrently, Brenner et al. (2026) find that transformers trained on constrained random walks factor their computation into a sufficient-statistic stage and a predictive-distribution stage, with the intermediate world-state representation shared across task variants.

Emergent world models and state tracking in sequence models. A parallel literature asks whether sequence models trained on surface tokens build models of the generating world. Li et al. (2023) recover board state from an Othello-playing transformer and Nanda et al. (2023) show that a closely related encoding, relative to the current player, is linear; Liu et al. (2023) characterize the depth at which transformers can track automaton state. Vafa et al. (2024) caution that models passing standard world-model diagnostics can nonetheless hold incoherent state, using metrics derived from the Myhill–Nerode theorem—i.e. from exactly the causal-equivalence relation of Appendix M.3. Pereira & Zuidema (2026) sharpen the caution further: models can encode a geometrically faithful world model yet fail to maintain it through a reasoning trace, with performance partially recovered by injecting the earlier representation at inference time. Our setting is the complementary one: rather than inferring an unknown world model from behavior, we specify the process, compute its belief geometry and dimension exactly, and ask whether the network finds them. The emergence of beliefs we study is orthogonal to approaches to deliberately instill beliefs in transformers by architectural choices (Hu et al., 2025; Teoh et al., 2025).

Probing, and the gap between decodability and use. Linear probes have been the standard instrument for representational claims since Alain & Bengio (2016), and their pitfalls are equally well documented: a probe's accuracy conflates what the representation encodes with what the probe itself can learn, motivating control tasks and selectivity (Hewitt & Liang, 2019) and, more broadly, the methodological cautions surveyed by Belinkov (2022). Elazar et al. (2021) sharpen the central objection—high decodability does not establish that the model *uses* the decoded quantity—which is precisely why our probe results (Fig. 2) are insufficient on their own and why we report the intervention of Fig. 3. We adopt both practices: selectivity controls in Fig. 11 and a causal edit measured against null, floor, and unrelated baselines.

Causal interventions and steering. Our belief-patching experiment inherits the activation-patching paradigm introduced for causal mediation analysis by Vig et al. (2020) and developed into causal tracing and targeted weight edits by Meng et al. (2022); we follow the metric and baseline recommendations of Zhang & Nanda (2024). Most directly comparable is distributed alignment search (Geiger et al., 2024), which localizes an interpretable causal variable to a learned subspace and intervenes there—analogous to our write along the decoder matrix, with the difference that our target subspace is given by theory rather than learned. The activation-steering literature (Turner et al., 2023; Rimsky et al., 2024; Marks & Tegmark, 2024) pursues behavioral control through similar edits; our results suggest that where a process factorizes, the corresponding edits are naturally *modular*, addressing one agent's beliefs without disturbing another's.

Factored, disentangled, and modular representations. The disentanglement literature (Higgins et al., 2017; 2018) confronts an identifiability barrier: Locatello et al. (2019) show that unsupervised disentanglement is impossible without inductive bias or supervision. Our known factorized generator gives a controlled setting and a theoretically derived target for testing factorization (Corollary 2). Work seeking modularity without a known ground truth—differentiable weight masks (Csordás et al., 2021), clusterability (Filan et al., 2021), and structural compositionality probes (Lepori et al., 2023)—is the empirical counterpart to our synthetic setup, where the factors are known in advance. Architecturally imposed modularity (Shazeer et al., 2017; Fedus et al., 2022) contrasts with the emergent modularity we observe.

Feedback loops, cybernetics, and predictive processing. That an adaptive system in a loop with its environment must match the variety of that environment is the founding insight of cybernetics (Ashby, 1956; 1960), sharpened into an internal-model claim by Conant & Ashby (1970) and later formalized by e.g. (Richens et al., 2025; Baltieri et al., 2025; Virgo et al., 2025): every good regulator of a system must be a model of that system. Our Corollary 2 is a formal refinement in the predictive-state setting—not merely that a model is required, but that in a feedback loop it decomposes into the environment's belief and the action-delayed agent's belief. Information-theoretic treatments of the perception–action cycle (Klyubin et al., 2005; Tishby & Polani, 2011; Still, 2009) quantify the channel between agent and world; the thermodynamic costs of running such loops and of modularity itself are treated by Boyd et al. (2016; 2018); Fiderer et al. (2025), and the decomposition of monolithic world models into modular sub-transducers by Boyd et al. (2025). In neuroscience, predictive processing posits a hierarchy of predictive subunits coupled through prediction error (Rao & Ballard, 1999; Friston, 2010; Clark, 2013), the biological instantiation of the modular-predictor picture that motivates our question about transformers.

Multi-agent systems and models of other agents. Minsky (1986) is the canonical statement that apparently monolithic cognition may be a society of interacting parts. In language models, Andreas

(2022) argues that next-token prediction induces representations of latent agents, and Zhu et al. (2024) find the beliefs of a story's protagonist and of an omniscient observer linearly decodable from attention-head activations; our contribution is to make the corresponding geometry and dimension exact rather than qualitative. Behavioral work probes whether models represent others' mental states (Sap et al., 2022; Kosinski, 2024) and deploys LLMs as populations of interacting agents (Park et al., 2023; Du et al., 2024). In multi-agent RL, agents that model each other in a loop are studied as opponent modeling (Foerster et al., 2018; Albrecht & Stone, 2018).

Abstractions, coarse-graining, and complexification. State abstractions are a central component of modern model-based RL Abel (2022); Subramanian et al. (2022); Ni et al. (2024); Zhang et al. (2026) used to understand the process of coarse-graining concrete states to abstract states while preserving decision-relevant behavioral properties, with connections to bisimulation learning Givan et al. (2003); Ravindran & Barto (2003); Zhang et al. (2019). Our approach here differs because we look at coarse-grainings of interfaces, rather than states.

That coarse-graining can *increase* the dimension needed to predict is counterintuitive but has precedent: functions of Markov chains are generally not Markov (Burke & Rosenblatt, 1958), and partial observation of a finite HMM can produce uncountably many or fractally distributed predictive states (Jurgens & Crutchfield, 2021b;a). Our contribution is to identify this mechanism in feedback settings, where factored structure is lost by summing over one factor's outputs, and the empirical finding that transformers track this increased complexity (Fig. 4e). This connects to the broader question of when macroscopic descriptions are more, rather than less, informative than microscopic ones (Anderson, 1972; Hoel et al., 2013; Rosas et al., 2020), and to the problem of where one agent ends and another begins (Krakauer et al., 2020; Friston, 2013; Bruineberg et al., 2022), which our framework raises but does not settle: any joint process admits *some* feedback decomposition, and Eq. (8) tells us only when a given decomposition pays for itself.

D JOINT PROCESS FOR A FEEDBACK LOOP

From the circuit shown in Fig. 1, we can construct the joint process over observables X and Y, and latents R and S by applying the transducers for agent T and agent U. They can be thought of as symbol-labeled transition matrices, with elements given by the conditional probabilities:

$$T^{(y|x)}_{r\to r'} = \Pr(Y_t = y, R_{t+1} = r' | X_t = x, R_t = r) \tag{22}$$

$$U^{(x|y)}_{s\to s'} = \Pr(X_{t+1} = x, S_{t+1} = s' | Y_t = y, S_t = s). \tag{23}$$

Given the perception-action loop, we must decide how to start. We might either choose Agent T or Agent U to go first. Based on our indexing, we choose to start with the initial joint probability $p_0(r, s, x) \equiv \Pr(R_0 = r, S_0 = s, X_0 = x)$ of Agent T's memory R, Agent U's memory S, and Agent U's action X. By applying agent T's transducer and agent U's transducer repeatedly, we obtain the joint process over all variables

$$\begin{aligned}\Pr(R_{0:L+1} = r_{0:L+1}, S_{0:L+1} = s_{0:L+1}, X_{0:L+1} = x_{0:L+1}, Y_{0:L} = y_{0:L}) \\ = p_0(r_0, s_0, x_0) \prod_{t=0}^{L-1} T^{(y_t|x_t)}_{r_t\to r_{t+1}} U^{(x_{t+1}|y_t)}_{s_t\to s_{t+1}}.\end{aligned} \tag{24}$$

The combination of the transducers T and U provides a Markov kernel for generating the process over both observables (X and Y) and latents (R and S).

However, as an external observer of the feedback process, the latents are exactly that, hidden from view. The external observer only sees the joint process over actions and perceptions $\Pr(X_{0:L}, Y_{0:L})$. We need only marginalize the latent trajectories in R and S to produce the joint process $\Pr(X_{0:L}, Y_{0:L})$

$$\Pr(X_{0:L} = x_{0:L}, Y_{0:L} = y_{0:L}) = \sum_{s_{0:L+1}, r_{0:L+1}, x_L} p_0(r_0, s_0, x_0) \prod_{t=0}^{L-1} T^{(y_t|x_t)}_{r_t\to r_{t+1}} U^{(x_{t+1}|y_t)}_{s_t\to s_{t+1}}, \tag{25}$$

This reveals the architecture of an HMM that uses the joint latent space $\mathcal{R}\otimes\mathcal{S}\otimes\mathcal{Z}$, where $\mathcal{Z}$ is a copy of the action space $\mathcal{X}$. Specifically, we introduce a new HMM for the joint process with elements

$$V^{(xy)}_{rsz\to r's'z'} = U^{(z'|y)}_{s\to s'} T^{(y|z)}_{r\to r'} \delta_{z,x}. \tag{26}$$

This arises by "boxing" the T and U circuit elements into a single V element as shown in Fig. 8. R_t, S_t, and X_t all carry information between elements, making them the new memory basis. It's also worth noting that (due to indexing) we copy X_t to the output of V using the delta function $\delta_{z,x}$.

We can reinterpret the linear operator for V as a tensor product, given the equation above. We begin by using the standard notation for an HMM linear operator

$$\hat{V}^{(xy)} \equiv \sum_{r,r'\in\mathcal{R}, s,s'\in\mathcal{S}, z,z'\in\mathcal{X}} |r'\rangle\langle r| \otimes |s'\rangle\langle s| \otimes |z'\rangle\langle z| V^{(xy)}_{rsz\to r's'z'}. \tag{27}$$

Then we apply the decomposition and find

$$\hat{V}^{(xy)} \equiv \sum_{r,r'\in\mathcal{R}, s,s'\in\mathcal{S}, z,z'\in\mathcal{X}} |r'\rangle\langle r| \otimes |s'\rangle\langle s| \otimes |z'\rangle\langle z| U^{(z'|y)}_{s\to s'} T^{(y|z)}_{r\to r'} \delta_{z,x} \tag{28}$$

$$= \sum_{z,z'\in\mathcal{X}} \left(\sum_{r,r'\in\mathcal{R}} T^{(y|z)}_{r\to r'} |r'\rangle\langle r| \right) \otimes \left(\sum_{s,s'\in\mathcal{S}} U^{(z'|y)}_{s\to s'} |s'\rangle\langle s| \right) \otimes \delta_{z,x} |z'\rangle\langle z| \tag{29}$$

$$= \sum_{z,z'\in\mathcal{X}} \hat{T}^{(y|z)} \otimes \hat{U}^{(z'|y)} \delta_{z,x} \otimes |z'\rangle\langle z| \tag{30}$$

$$= \sum_{z'\in\mathcal{X}} \hat{T}^{(y|x)} \otimes \hat{U}^{(z'|y)} \otimes |z'\rangle\langle x| \tag{31}$$

$$= \hat{T}^{(y|x)} \otimes \sum_{z'\in\mathcal{X}} \hat{U}^{(z'|y)} \otimes |z'\rangle\langle x|. \tag{32}$$

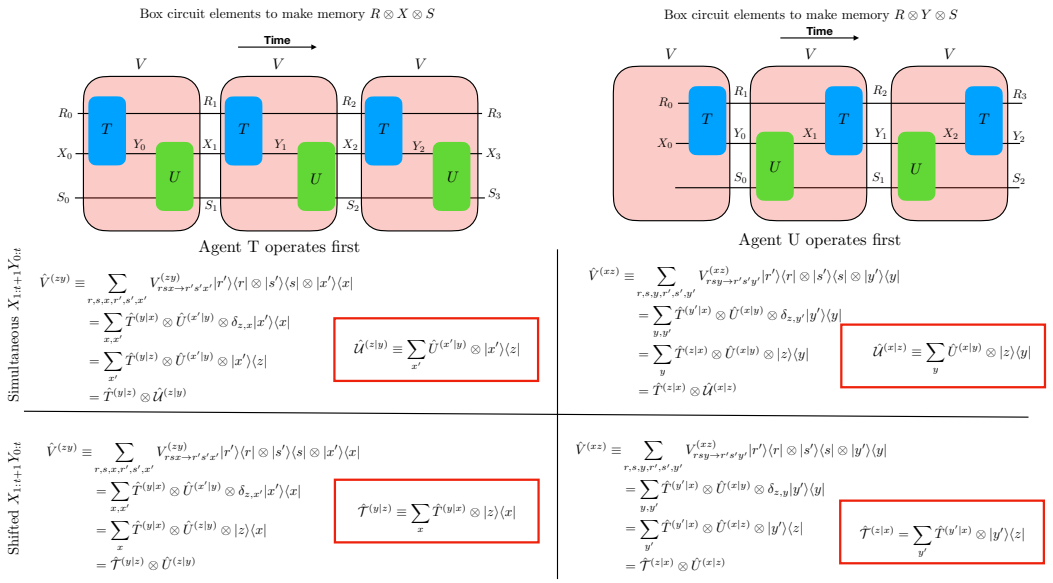


Figure 8: There are four different readily accessible presentations for a feedback process between agent T and U. Which one you arrive at depends on 1) whether you choose to model the simultaneous process $X_{0:t}Y_{0:t}$ or the shifted process $X_{1:t+1}Y_{0:t}$, and 2) whether you box circuit elements so that Agent U or Agent T operates first. All presentations modify the linear operator of one of the transducers, either delaying the action or the perception.

We introduce a new operator on the joint space of U's actions and memory

$$\hat{\mathcal{U}}^{(x|y)} \equiv \sum_{z' \in \mathcal{X}} \hat{U}^{(z'|y)} \otimes |z'\rangle\langle x|, \tag{33}$$

which is the action-delayed agent, because it returns the probability of U's last action. The joint HMM linear operator is the tensor product of agent T's linear operator with action-delayed U's linear operator:

$$\hat{V}^{(xy)} = \hat{T}^{(y|x)} \otimes \hat{\mathcal{U}}^{(x|y)}. \tag{34}$$

It should be noted that this particular decomposition is the result of broken symmetries in how we analyze the joint process. For instance, we might as easily try to predict the $X_{t+1}Y_t$ from its past as X_tY_t. This corresponds to predicting the shifted process $\Pr(X_{1:t+1}, Y_{0:t})$ rather than the simultaneous process $\Pr(X_{0:t}, Y_{0:t})$. Additionally, in boxing the joint feedback process, we can choose whether Agent T or Agent U operates first. This leads to four different valid presentations, as shown in Fig. 8.

E Proof that belief state updating retains the tensor product structure

The belief state update rule $|\rho_{t+1}\rangle = \frac{\hat{T}^{(x_t)}|\rho_t\rangle}{\langle 1|\hat{T}^{(x_t)}|\rho_t\rangle}$ preserves product structure:

$$
\begin{aligned}
&|\rho_{t+1}\rangle_{R,S,X} \\
&= \frac{\hat{V}^{(x_t y_t)}|\rho_t\rangle_{R,S,X}}{\langle 1|_{R,S,X}\hat{V}^{(x_t y_t)}|\rho_t\rangle_{R,S,X}} \\
&= \frac{\left(\hat{T}^{(y|x)} \otimes \hat{\mathcal{U}}^{(x|y)}\right)\left(|\rho_t\rangle_R \otimes |\rho_t\rangle_{S,X}\right)}{\left(\langle 1|_R \otimes \langle 1|_{S,X}\right)\left(\hat{T}^{(y|x)} \otimes \hat{\mathcal{U}}^{(x|y)}\right)\left(|\rho_t\rangle_R \otimes |\rho_t\rangle_{S,X}\right)} \\
&= \frac{\hat{T}^{(y|x)}|\rho_t\rangle_R \otimes \hat{\mathcal{U}}^{(x|y)}|\rho_t\rangle_{S,X}}{\langle 1|_R\hat{T}^{(y|x)}|\rho_t\rangle_R\langle 1|_{S,X}\hat{\mathcal{U}}^{(x|y)}|\rho_t\rangle_{S,X}} \\
&= \frac{\hat{T}^{(y|x)}|\rho_t\rangle_R}{\langle 1|_R\hat{T}^{(y|x)}|\rho_t\rangle_R} \otimes \frac{\hat{\mathcal{U}}^{(x|y)}|\rho_t\rangle_{S,X}}{\langle 1|_{S,X}\hat{\mathcal{U}}^{(x|y)}|\rho_t\rangle_{S,X}} \\
&= |\rho_{t+1}\rangle_R \otimes |\rho_{t+1}\rangle_{S,X}.
\end{aligned}
\tag{35}
$$

F Factored Social Network Operator

The feedback loop HMM in a social network can be expressed as

$$
\hat{V}^{(\vec{x}\vec{y})} = \hat{H}^{(\vec{y}|\vec{x})} \otimes \hat{\vec{\mathcal{A}}}^{(\vec{x}|\vec{y})}, \tag{36}
$$

where $\hat{H}^{(\vec{y}|\vec{x})}$ is the social network hub and the action-delayed macro-agent is

$$
\hat{\vec{\mathcal{A}}}^{(\vec{x}|\vec{y})} \equiv \sum_{\vec{z}\in\vec{\mathcal{X}}} \hat{\vec{A}}^{(\vec{z}|\vec{y})} \otimes |\vec{z}\rangle\langle\vec{x}|. \tag{37}
$$

The macro-agent is composed of conditionally independent micro-agents, so it can be expressed as a tensor product

$$
\hat{\vec{A}}^{(\vec{x}|\vec{y})} = \bigotimes_{i=0}^{N-1} A(i)^{(x(i)|y(i))}. \tag{38}
$$

This, in turn, yields an expression for the action-delayed macro-agent as a tensor product of the action-delayed micro-agents:

$$
\hat{\vec{\mathcal{A}}}^{(\vec{x}|\vec{y})} = \bigotimes_{i=0}^{N-1} \hat{\mathcal{A}}(i)^{(x(i)|y(i))}, \tag{39}
$$

where the action-delayed micro-agent is expressed

$$
\hat{\mathcal{A}}(i)^{(x(i)|y(i))} = \sum_{z(i)\in\mathcal{X}(i)} \hat{A}(i)^{(z(i)|y(i))} \otimes |z(i)\rangle\langle x(i)|. \tag{40}
$$

G Decomposing Joint Processes into Feedback Agents

Any joint process can be decomposed into two transducers interacting in a feedback loop. To illustrate this, let us consider a joint process

$$
\Pr(X_{0:L}, Y_{0:L}). \tag{41}
$$

We can build this up from past to future using the probability of the present conditioned on the past:

$$
\Pr(X_{0:L}, Y_{0:L}) = \prod_{t=0}^{L-1} \Pr(X_t, Y_t | X_{0:t}, Y_{0:t}). \tag{42}
$$

Each of these conditional probabilities can be decomposed using Bayes' rule:

$$\Pr(X_t, Y_t | X_{0:t}, Y_{0:t}) = \Pr(X_t | Y_t, X_{0:t}, Y_{0:t}) \Pr(Y_t | X_{0:t}, Y_{0:t}). \tag{43}$$

Here we have indicated, through conditioning, that Y_t comes before X_t in the feedback loop. This is an arbitrary choice, and this proof can proceed as easily if the alternate time-indexing is used, where X_t precedes Y_t.

The decomposition of this process into an update of next Y then next X allows us to identify kernels that build the process step-by-step

$$k(X_t | X_{0:t}, Y_{0:t+1}) \equiv \Pr(X_t | Y_t, X_{0:t}, Y_{0:t}) \tag{44}$$
$$m(Y_t | X_{0:t}, Y_{0:t}) \equiv \Pr(Y_t | X_{0:t}, Y_{0:t}). \tag{45}$$

Each of these kernels can be used to generate an interface:

$$k(X_{0:L} | Y_{0:L}) \equiv \prod_{t=0}^{L-1} k(X_t | X_{0:t}, Y_{0:t+1}) \tag{46}$$
$$m(Y_{0:L} | X_{0:L}) \equiv \prod_{t=0}^{L-1} m(Y_t | X_{0:t}, Y_{0:t}), \tag{47}$$

Note that while the interfaces describe a stochastic channel from X to Y or vice versa, they are not necessarily equal to the true conditional probability between the two sequences:

$$k(X_{0:L} | Y_{0:L}) \neq \Pr(X_{0:L} | Y_{0:L}) \tag{48}$$
$$m(Y_{0:L} | X_{0:L}) \neq \Pr(Y_{0:L} | X_{0:L}). \tag{49}$$

Instead, their product is the joint process:

$$\Pr(X_{0:L}, Y_{0:L}) = k(X_{0:L} | Y_{0:L}) m(Y_{0:L} | X_{0:L}). \tag{50}$$

This factored form has been examined before in the context of feedback Massey (1990); Tatikonda & Mitter (2009). Our contribution stems from recognizing that *both* interfaces can be realized by transducers linked into this feedback loop.

If these interfaces are causal (nonanticipatory) (Boyd et al., 2025), then this is precisely the form of the joint probability when two agents interact in a feedback loop, otherwise known as a perception-action loop in Fiderer et al. (2025). In this case, there exist transducer presentations for both interfaces, and they can produce the observed joint process if linked into a feedback loop.

To complete the proof, we need simply verify that these interfaces are nonanticipatory. This means that future inputs cannot influence past outputs

$$k(X_{\text{past}} | Y_{\text{past}}, Y_{\text{future}}) = k(X_{\text{past}} | Y_{\text{past}}) \tag{51}$$
$$m(Y_{\text{past}} | X_{\text{past}}, X_{\text{future}}) = m(Y_{\text{past}} | X_{\text{past}}). \tag{52}$$

The past/future partitioning is asynchronous in a feedback loop. In this case, the past inputs for k are $Y_{0:t}$, because Y_t immediately precedes X_t, while the past inputs for m are $X_{0:t-1}$, because X_{t-1} immediately precedes Y_t. Therefore, causality in this case reduces to two relations for all t:

$$k(X_{0:t} | Y_{0:L}) = k(X_{0:t} | Y_{0:t}) \tag{53}$$
$$m(Y_{0:t} | X_{0:L}) = m(Y_{0:t} | X_{0:t-1}). \tag{54}$$

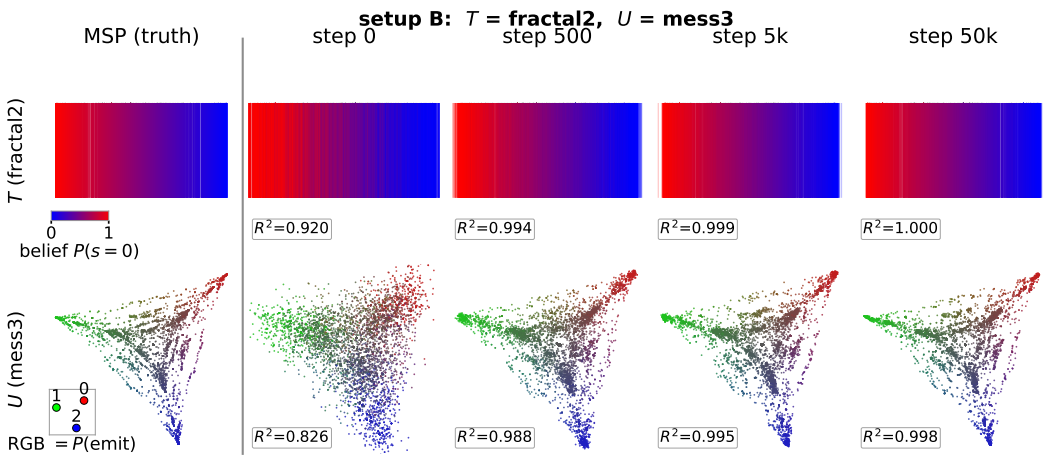


Figure 9: **Belief-state geometry of the role-swapped setup likewise refines over training.** Rows: both factors of setup B (agent T fractal2, agent U mess3); setup A is shown in Fig. 2. Leftmost column: ground-truth MSP. Remaining columns: linear reconstruction from the final-layer residual stream at checkpoints 0, 500, 5k, and 50k (probe refit per checkpoint; same PCA basis and color conventions as Fig. 2; held-out R^2 per panel). Both factors are already reconstructed to $R^2 \geq 0.83$ at initialization and are refined by training to $R^2 \geq 0.998$ by step 50k.

Finally, we can employ marginalization to validate the above terms:

$$k(X_{0:t}|Y_{0:L}) \equiv \sum_{X_{t:L}} k(X_{0:L}|Y_{0:L}) \tag{55}$$

$$= \sum_{X_{t:L}} \prod_{i=0}^{L-1} k(X_i|X_{0:i}, Y_{0:i+1}) \tag{56}$$

$$= \left(\prod_{i=t}^{L-1} \sum_{X_i} k(X_i|X_{0:i}, Y_{0:i+1})\right) \left(\prod_{i=0}^{t-1} k(X_i|X_{0:i}, Y_{0:i+1})\right) \tag{57}$$

$$= k(X_{0:t}|Y_{0:t}) \tag{58}$$

$$m(Y_{0:t}|X_{0:L}) \equiv \sum_{Y_{t:L}} m(Y_{0:L}|X_{0:L}) \tag{59}$$

$$= \sum_{Y_{t:L}} \prod_{i=0}^{L-1} m(Y_i|X_{0:i}, Y_{0:i}) \tag{60}$$

$$= \left(\prod_{i=t}^{L-1} \sum_{Y_i} m(Y_i|X_{0:i}, Y_{0:i})\right) \left(\prod_{i=0}^{t-1} m(Y_i|X_{0:i}, Y_{0:i})\right) \tag{61}$$

$$= m(Y_{0:t}|X_{0:t-1}). \tag{62}$$

Thus, the interfaces k and m are non-anticipatory, and can each be generated by a transducer. Therefore, the joint process can be generated by those two respective transducers in a feedback loop. □

H Additional evidence for belief geometry

Fig. 2 shows a feedback loop between a mess3 transducer as agent T and a fractal2 transducer as U. To form a predictive state of this process, the latent state of the succeeding factor U is augmented by the additional time-latched action X. What happens when the two factors are swapped? The token sequence that this process describes is still the same as before, with the crucial difference that the tuples of tokens which are predicted together for the same timestep are now offset by one. For this

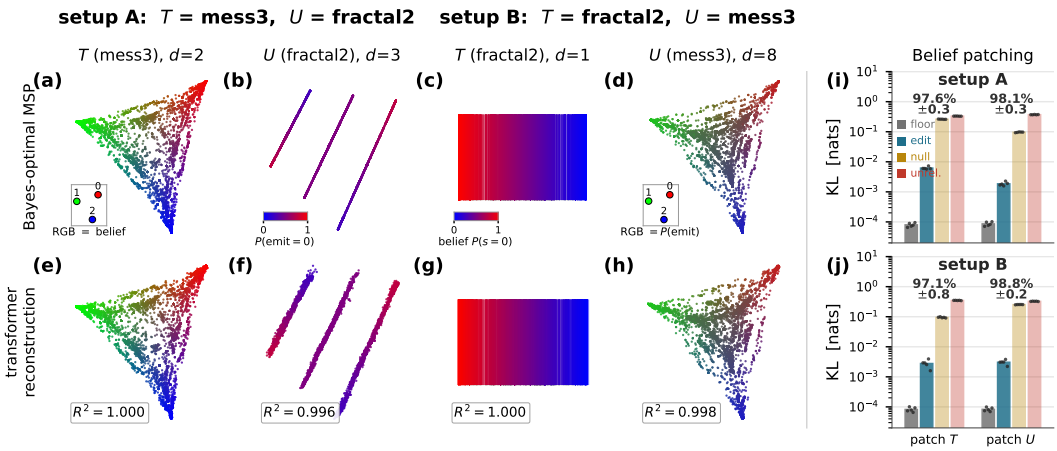


Figure 10: **Both mess3–fractal2 setups are represented by factorized, causally load-bearing beliefs in the residual stream.** The final-checkpoint reconstruction of Fig. 2 with the role-swapped arrangement alongside it: mess3 as agent T and fractal2 as agent U (setup A, panels a,b,e,f — the setup reproduced in the main text) or vice-versa (setup B, panels c,d,g,h). Top row: Bayes-optimal MSP; bottom row: reconstruction from the final-layer residual stream at step 50k. Panel and color conventions and the probe protocol are as in Fig. 2; the belief patching in **(i,j)** is as in Fig. 3.

reason, now the action of the mess3 process needs to be latched for the prediction of the next token tuple, which extends the latent space of the mess3 factor. In both the original setup (setup A) and the role-swapped setup (setup B), the beliefs in the residual stream are refined over training (see Figs. 2 and 9) to $R^2 \geq 0.996$, and patching either factor closes 97–99% of the null-to-target distance (see Fig. 10).

We performed two additional controls to verify these results. First, we show that the probe error of the true belief is significantly lower than a shuffled or a time-shifted control over all steps of training (see Fig. 11a,b). In addition, we tested probing from different layers to find that later network layers have lower probe errors in our experiments (see Fig. 11c-f), justifying our choice of analyzing the last layer in our main figures.

I THE FRACTAL2 TELEPHONE PROCESS

Fig. 4c reports the cumulative explained variance for the three-agent mess3 telephone process. Fig. 12 is the same measurement for five fractal2 sub-agents. Here the residual stream is *lower*-dimensional than even the factored reference: fractal2's mixed-state presentation occupies two one-dimensional sub-manifolds of its probability simplex (Fig. 10c), so the factored concatenation is itself a loose bound on what the network has to represent.

J EFFECTIVE RANK ACROSS MODEL WIDTHS

To ensure our results are robust to changes in the transformer architecture, we ran additional variations in which we changed the residual stream width. Sweeping the residual stream width from 60 to 240 while keeping the process fixed, we find that the effective rank of the residual stream is largely independent of the architecture (see Fig. 13). The number of components at the 80% and 95% CEV thresholds remains consistent across widths, indicating that the effective rank is determined by the process rather than the model's architecture. This is further supported by the fact that all runs at different widths yield an $n_c@80\%$ of 5, which is below the factored bound of 24 and well below the joint-Kronecker bound of 728.

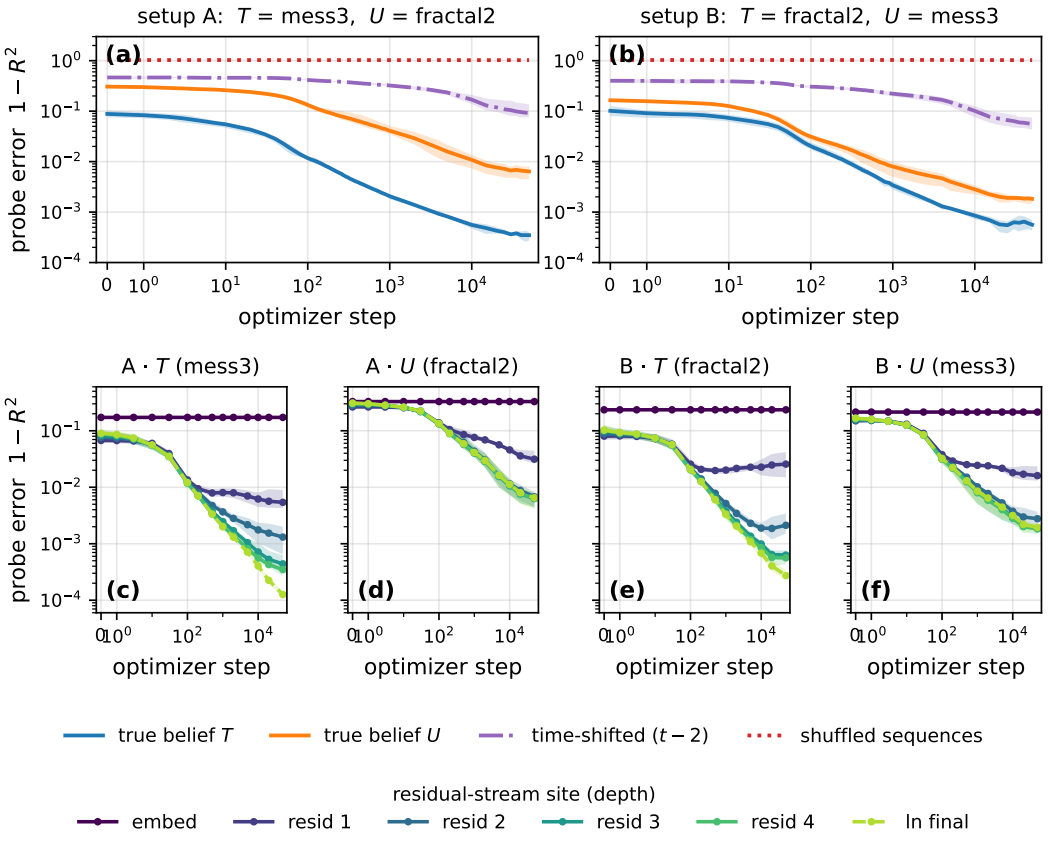


Figure 11: **Probe selectivity controls and layer-wise emergence.** All panels show the held-out linear-probe error $1 - R^2$ on the residual stream over training (mess3–fractal2 feedback loops). Solid curves: mean over five seeds per setup — the canonical run plus four reruns (Table 2); shaded bands: min–max across seeds. **(a,b)** Final-layer probes fitted to the true per-factor beliefs and to two controls: the same sequence's belief two positions earlier (time-shifted) and beliefs of shuffled sequences (a derangement of the batch). The shuffled floor stays at $1 - R^2 \approx 1$ throughout; the true factors improve by 40–280$\times$ over training. **(c–f)** The true-belief probe refit at each residual-stream depth (embedding, blocks 1–4, final LayerNorm). Blocks 1–2 acquire essentially all of the structure; the action-delayed factor U is complete by block 2, while factor T sharpens to the last layer. The embedding row is training-invariant by construction (a linear probe on it can realize any function of token and position).

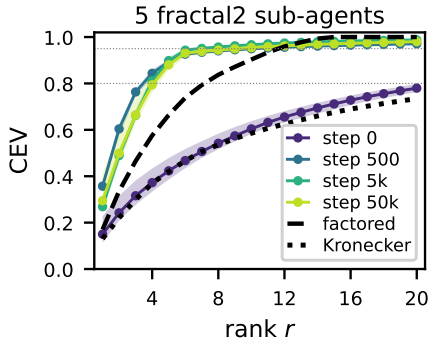


Figure 12: **Cumulative explained variance for the five-agent fractal2 telephone process.** Same protocol and conventions as Fig. 4c: CEV of the final-layer residual stream by rank at several training steps; solid curves the mean over five seeds, shaded bands the min–max across them; dashed the CEV of the factored concatenation of the Bayes-optimal beliefs, dotted that of their Kronecker product.

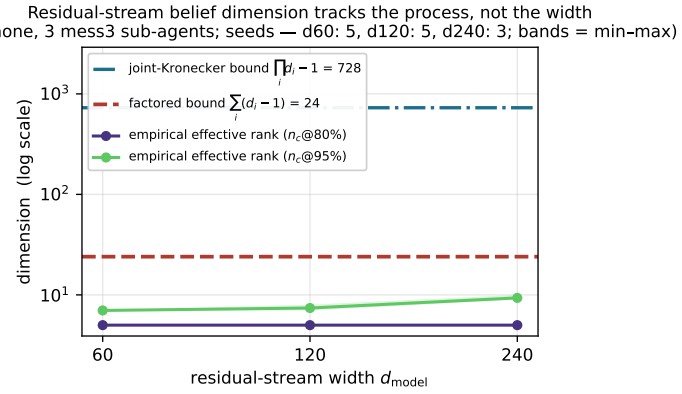


Figure 13: **Effective rank is mostly independent of model width.** Effective rank of the final-layer residual stream at step 50000 (BOS position excluded, 1024 sequences, same protocol as Fig. 4d) for the three-agent mess3 telephone process, against residual-stream width $d_{\text{model}} \in \{60, 120, 240\}$: n_c at the 80% and 95% CEV thresholds. Points are means over five seeds for widths 60 and 120, and three seeds for width 240. bands are min–max across seeds and are degenerate for n_c@80%. Horizontal lines: the factored bound $\sum_i (d_i - 1) = 24$ (dashed) and the joint-Kronecker bound $\prod_i d_i - 1 = 728$ (dash-dotted).

K Simulated processes and parameters

The transducers in our simulations are built from three base HMMs, mess3, fractal2 and sns (see Figure 14), with the input tokens selecting one of multiple parametrizations of the respective HMM (see Table 1).

If not specified otherwise, the same four-layer transformer with a residual stream width of 120 and three heads with a head dimension of 40 has been used (see Table 2).

The models are trained online for $50,000$ training steps and converge close to the Bayes-optimal floor: across all five seeds per setup, runs converge to within a few 10^{-3} nats of their floors in all cases except two of the five $N{=}4$ seeds, which settle 3–5 $\times$ 10^{-3} nats above it (see Fig. 15).

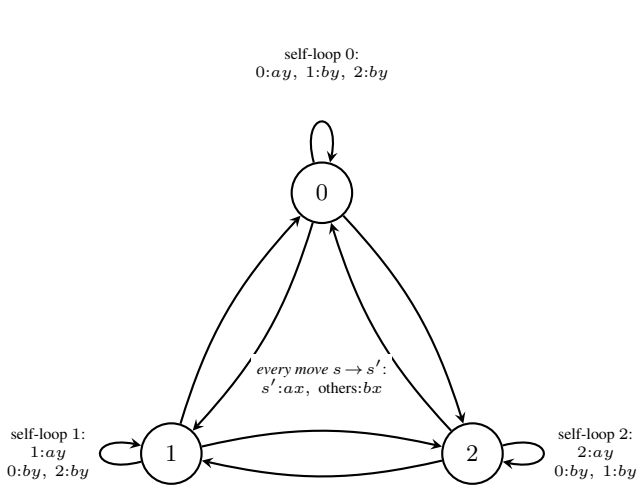
For each figure, a fixed batch of 1024 sequences is sampled from the generative process together with the Bayes-optimal per-factor belief states (the MSP coordinates) along each sequence. *MSP panels*: the ground-truth beliefs of a factor are projected onto their own top-two PCA directions; the model reconstruction is an ordinary-least-squares linear probe from the final-layer residual stream (`resid_post`, BOS position excluded) to the belief coordinates, fit on 80% of sequences, evaluated on the held-out 20% (reported R^2), and projected onto the *same* PCA basis as the ground truth. For two-state factors the belief simplex is one-dimensional and PCA direction two is degenerate; those panels are drawn as barcode plots, with one vertical line per sample at its belief coordinate. *CEV panels*: cumulative explained variance by rank of the centered final-layer residual stream (SVD), per checkpoint, compared against the CEV of the concatenated per-factor beliefs (factored reference) and of their Kronecker product (joint reference) on the same batch.

The code to reproduce all figures is based on the implementation used by Shai et al. (2026), and is added as supplementary material for anonymity, and will be published together with the final version of this manuscript.

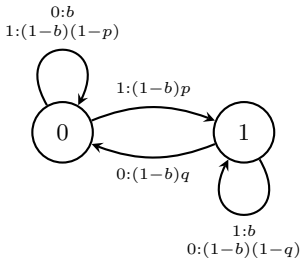
L Interventional test of belief factorization

Linear decodability of per-factor beliefs (Fig. 2) is correlational. Here we test the *causal* counterpart: does the trained network actually use the observed factored representations to predict the next token? To test this question, we patch counterfactual beliefs into a single factor's residual-stream subspace and check whether the output distribution shifts exactly as the Bayes-optimal predictor would if only that factor's belief had changed.

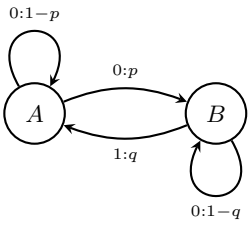
For a fixed batch of 1024 sequences with exact per-factor Bayes beliefs b_f, we collect the final-layer residual stream h (`resid_post`, before the final LayerNorm) at all non-BOS positions and fit, by



(a) mess3(x, a)



(b) fractal2(b, p, q)



(c) sns(p, q)

Figure 14: **The base HMMs** (edge labels read "emission : probability"). **(a)** mess3(x, a): three states, three symbols; with $b = (1 - a)/2$ and $y = 1 - 2x$. A transition into state s' emits symbol s' with probability $ax \equiv ax$ and each other symbol with probability bx (independent of the source state, so both incoming arrows share the label shown by each vertex); a self-loop at s emits s with probability ay and each other symbol with probability by. Its mixed-state presentation is the fractal simplex of Fig. 2 (top row). **(b)** fractal2(b, p, q): two states, two symbols; state s re-emits its own symbol with probability b and stays, otherwise emits the other symbol and moves with probability p (resp. q). Its two-state belief simplex is the barcode of Fig. 10c. **(c)** sns(p, q): the simple nonunifilar source; symbol 1 is emitted only on the resetting $B \to A$ transition, giving a Cantor-like 1-D belief geometry, drawn as a barcode like the fractal2 factor of Fig. 10c.

ordinary least squares, a single *joint decoder*

$$h \;\approx\; c + \textstyle\sum_f E_f\, b_f, \tag{63}$$

whose blocks E_f serve as per-factor *write* matrices, together with per-factor *read* probes $b_f \approx W_f^\top h + a_f$. (For the telephone process, the trivial router factor is omitted.) Fitting this joint decoder instead of inverting the belief probes of the correlational results has been crucial for successful patching, as probes appear to sometimes fit smaller variance directions, which mirror the belief state but have little causal power.

Counterfactual beliefs come from a random donor permutation π of the batch, $b'_f = b_{f,\pi(i)}$, taken at the same sequence position (the spliced tuple (b'_f, b_{-f}) is in general *not* jointly realizable). At the last trained position we add

$$\Delta \;=\; E_f\,(b'_f - b_f), \qquad \Delta_\perp \;=\; \Delta - \big(\Delta\, E_{-f}\big)\, E_{-f}^{+}, \tag{64}$$

where E_{-f} stacks the write matrices of all other factors and $(\cdot)^+$ is the pseudoinverse: the edit is orthogonalized against the subspace in which the residual stream *encodes* the other factors' beliefs, so the write cannot directly alter any partner's belief encoding. Projecting out the partners' read probes W_{-f} instead would be nearly vacuous: the raw write already lies almost entirely in the probes' null space (the projection removes 0.1–0.3% of its norm and changes no reported number relative to using no projection at all), because a factor's read-probe and encoder subspaces are themselves nearly orthogonal (mean overlap 0.05 across setups). The encoder projection removes 23–66% of the write. Neither construction is uniformly conservative: the encoder projection also removes whatever part of the *intended* write falls in the overlap of the encoder spans, deflating steering rates where that overlap is large, while the unprojected write can corrupt the partners' encodings directly, deflating

Table 1: **Generative-process parameters.** Sequences have length 8 plus a BOS token. Each agent's kernels are stacked along its input axis: the incoming symbol selects the kernel. Agent U is the action-delayed factor, whose state is augmented by the latched action ($\times A$). Setup A is the mess3–fractal2 loop shown in Fig. 2, and the pair (A, B) is shown in Fig. 10. The primed setups (A′, B′) replace fractal2 by sns; they are the mess3–sns pair of Appendix L. In both pairs, setup A has mess3 as agent T. The mess3↔mess3 loop appears in Fig. 16. The fractal2 kernels use $p = q$: fractal2's state-transition matrix is doubly stochastic exactly then, so the uniform initial belief used by the belief machinery is the true stationary distribution.

setup	factor	kernels (parameters)
feedback A (vocab 6)	agent T = mess3 (S=3) agent U = fractal2 (S=2, A=2)	$(x, a) \in \{(0.05, 0.85), (0.15, 0.60)\}$ $(b; p{=}q) \in \{(0.25; 0.25), (0.50; 0.35), (0.75; 0.65)\}$
feedback B (vocab 2·3=6)	agent T = fractal2 (S=2) agent U = mess3 (S=3, $\times A$=3)	$(b; p{=}q) \in \{(0.25; 0.25), (0.50; 0.35), (0.75; 0.65)\}$ $(x, a) \in \{(0.05, 0.85), (0.15, 0.60)\}$
feedback A′ (vocab 6)	agent T = mess3 (S=3) agent U = sns (S=2, $\times A$=2)	$(x, a) \in \{(0.05, 0.85), (0.15, 0.60)\}$ $(p, q) \in \{(0.30, 0.50), (0.50, 0.50), (0.70, 0.50)\}$
feedback B′ (vocab 2·3=6)	agent T = sns (S=2) agent U = mess3 (S=3, $\times A$=3)	$(p, q) \in \{(0.30, 0.50), (0.50, 0.50), (0.70, 0.50)\}$ $(x, a) \in \{(0.05, 0.85), (0.15, 0.60)\}$
loop mess3↔mess3 (vocab 3·3=9)	agent T = mess3 (S=3) agent U = mess3 (S=3, $\times A$=3)	$(x, a) \in \{(0.05, 0.60), (0.10, 0.50), (0.15, 0.70)\}$ $(x, a) \in \{(0.05, 0.60), (0.10, 0.50), (0.15, 0.70)\}$
telephone, fractal2 (N=5, A=2, vocab 4^5=1024)	agents 1–5	$(b; p{=}q)$ pairs, kernel 0 / kernel 1: A_1: (0.25; 0.25) / (0.75; 0.75) A_2: (0.30; 0.30) / (0.70; 0.70) A_3: (0.20; 0.30) / (0.80; 0.70) A_4: (0.35; 0.325) / (0.65; 0.675) A_5: (0.40; 0.30) / (0.60; 0.70)
telephone, mess3 ($N \in \{2, \ldots, 5\}$, A=3, vocab 9^N)	agent i, kernels	(x, a) triples (percept digit 0/1/2): A_1: (0.05, 0.85) (0.10, 0.50) (0.15, 0.70) A_2: (0.05, 0.60) (0.10, 0.50) (0.15, 0.70) A_3: (0.03, 0.80) (0.12, 0.55) (0.15, 0.65) A_4: (0.06, 0.75) (0.09, 0.45) (0.14, 0.60) A_5: (0.04, 0.70) (0.11, 0.65) (0.13, 0.55)

shielding for reasons that are artifacts of the write rather than properties of the network. The main text reports the unprojected write Δ; Fig. 16 compares both constructions across every setup, together with the between-factor decoder subspace overlap that determines how much the projection removes. In every setup that steers, the two constructions agree to within a few percentage points. The model is then run forward from the edited activation, and its next-token distribution (renormalized over the process vocabulary, excluding BOS) is compared against *exact* Bayes references computed from the generative process by evaluating the observation distribution on spliced belief tuples.

The edited output is compared, in KL divergence on the joint next-token distribution, against *exact* Bayes references evaluated on patched belief tuples: $H_0 = P(\cdot \mid b_T, b_U)$ (edit inert), $H_1 = P(\cdot \mid b'_f, b_{-f})$ (the factorized single-factor prediction — the patching target), and $H_2 = P(\cdot \mid b'_T, b'_U)$ (a wholly unrelated sequence). The headline quantity is the *closed gap* $1 - \mathrm{KL}(p_{\text{edited}} \,\|\, H_1)/\mathrm{KL}(p_{\text{clean}} \,\|\, H_1)$ — the fraction of the distance between an inert write and the patching target that the edit actually covers — reported per patched factor as a mean over evaluable interventions. Interventions whose references are closer than 0.05 total variation are excluded as unidentifiable ($\sim$70–78% of the batch remains).

Figure 16a summarizes belief patching across every setup. All setups without an sns factor close 96–99% of the null-to-target gap under the unprojected write (94–99% under the orthogonalized one; the largest single change is the first sub-agent of the mess3 telephone, $98\% \to 90\%$, which is also where the projection removes the most write mass, 52%). The two sns-containing loops, however, deviate from this finding: the closed gaps drop to 74%/90% (setup A′, patching mess3 resp. sns) and

Table 2: **Transformer and optimization hyperparameters**, shared by all runs (exceptions noted).

parameter	value
layers / d_{model} / heads / d_{head} / d_{mlp}	4 / 120 / 3 / 40 / 480
activation / normalization / positions	ReLU / LayerNorm / learned absolute
context n_{ctx}	9 (BOS + 8 tokens)
vocabulary	joint process vocab + 1 (BOS)
optimizer	AdamW, lr 5×10^{-4}, $\beta = (0.9, 0.999)$, weight decay 0
LR schedule	windowed reduce-on-plateau (factor 0.7, patience 10^4 steps, cooldown 2×10^4 steps, min lr 10^{-7})
batch size	25000 sequences, sampled fresh each step (online); 8192 for telephone mess3 N=5 (memory)
training steps	50000
seeds (process / weight init)	canonical runs: 77 / 7 (canonical A: 1102 / 1152, B: 1101 / 1151); four additional seeds per setup: feedback A 1301–1304, feedback B 1001–1004, feedback A′ 201–204, feedback B′ 101–104, loop mess3↔mess3 901–904, telephone fractal2 801–804, telephone mess3 N=2: 501–504, N=3: 401–404, N=4: 601–604, N=5: 701–704 (weight-init seed = process seed + 50; data streams of the reruns and of canonical A/B are seeded per run, the seed-77 canonical runs share the legacy step-indexed stream)
width sweep (Appendix J)	$d_{\text{model}} \in \{60, 240\}$ with 3 heads, $d_{\text{head}} = d_{\text{model}}/3$, $d_{\text{mlp}} = 4\,d_{\text{model}}$; seeds 2001–2005 resp. 2101–2105 ($d_{\text{model}} = 120$ is the canonical telephone mess3 family)
evaluation	every 10^3 steps on fresh sequences; exact Bayes floor from 32768 MC samples
checkpoints	log-spaced schedule (∼78 per run)

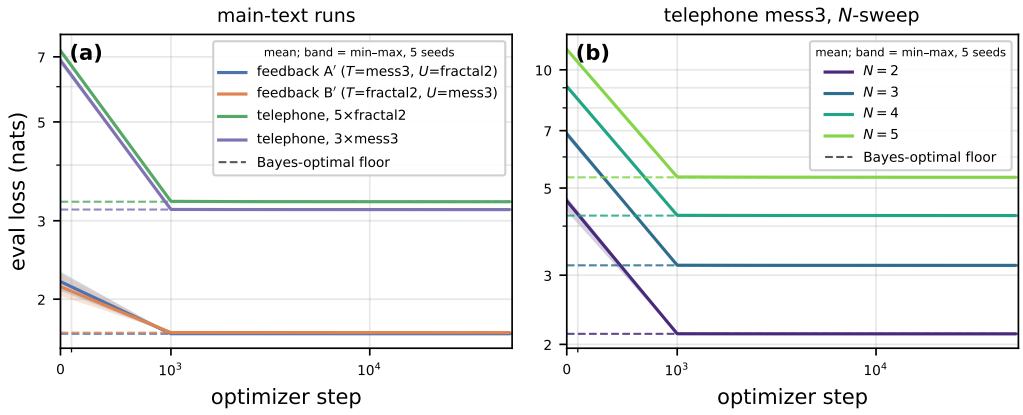


Figure 15: **All runs converge to their Bayes-optimal floors.** Evaluation loss over training (log scale); solid curves are means over five seeds per setup, shaded bands the min–max across seeds (mostly narrower than the line). Dashed lines: each setup's Bayes-optimal floor, position-averaged and estimated with 32768 MC samples (per-seed floor estimates differ by $\sim10^{-3}$ nats; the line shows the family mean). (a) The four main-text setups: the mess3–fractal2 feedback loops A and B of Figs. 2 and 10 and the two telephone processes. (b) The telephone-mess3 N-sweep of Fig. 4d. Every setup reaches its floor within $\sim10^3$ steps; final losses lie within 1–5 $\times$ 10^{-3} nats of the floor, i.e. at the precision of the MC estimate for all runs except two of the five N=4 seeds, which plateau 3–5 $\times$ 10^{-3} nats above it. The x-axis is symlog (linear below 10^3 to show the step-0 evaluation, logarithmic above).

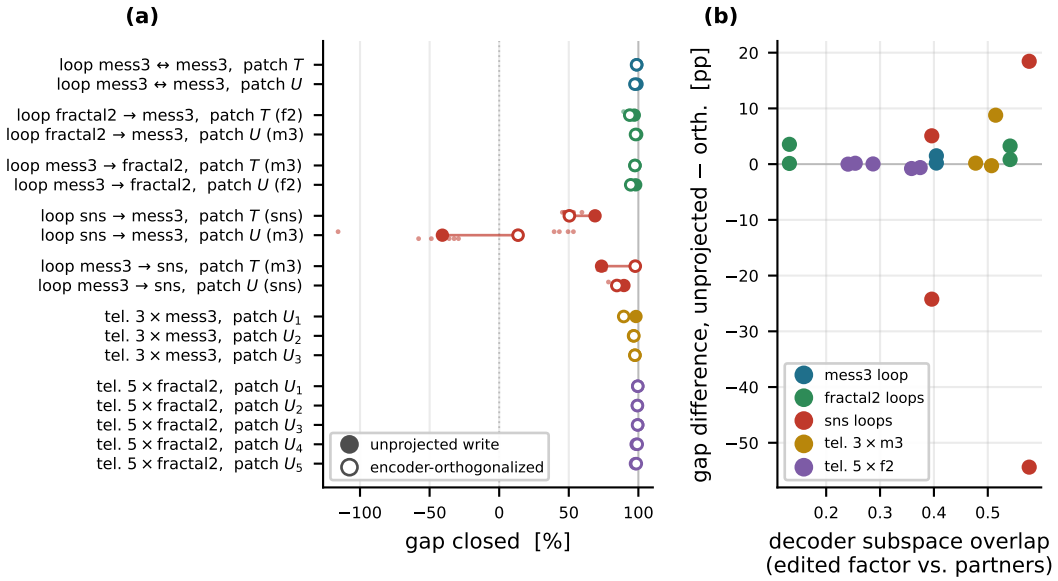


Figure 16: **Belief patching under the unprojected and the encoder-orthogonalized write. (a)** Fraction of the null-to-target distance closed by the edit, per setup and patched factor (large markers: ratio of means over seeds; small dots: individual seeds; filled: unprojected write Δ, open: $\Delta_\perp$ of Eq. (64)). Every setup without an sns factor closes 94–99% of the gap under either construction. The two sns loops fail under both, and are the only setups the construction changes qualitatively: under the unprojected write, editing the mess3 partner in the sns → mess3 loop leaves the output *farther* from the target than no edit at all (-41%), which the projection turns into $+14\%$ — the raw write was corrupting the sns encoding directly. **(b)** The between-factor decoder subspace overlap (share of the edited factor's encoder span lying in its partners') against the gap difference between the constructions. The overlap is substantial everywhere (0.13–0.58) and does not separate fragile from robust writes — the sns loops are the outliers in both directions — so the deviations are not explained by overlapping belief encodings.

$69\%/-41\%$ (setup B′, patching sns resp. its mess3 partner), while no other setup falls below 96%. Why this failure occurs remains a subject of future research.

Dose–response. The endpoint edit ($b_f \to b'_f$) leaves open whether the write works only between realizable belief states. We therefore interpolate the edit, splicing $b_f(\lambda) = (1-\lambda)\, b_f + \lambda\, b'_f$ for $\lambda \in \{0, 0.125, \ldots, 1\}$ — a valid simplex point, but in general *not* a reachable mixed state of the process — and compare the edited output against the Bayes prediction at the interpolated tuple $(b_f(\lambda), b_{-f})$. Figure 17 shows the result for the two mess3–fractal2 loops (five seeds each, unprojected write): at intermediate dose, the edited output lies one to two orders of magnitude closer to the interpolated target than to either endpoint reference — the endpoint references cross near $\lambda = 0.5$ while the on-path curve passes beneath both — and the edit closes 81–99% of the inert-write-to-target gap at every λ. These results suggest that the transformer truly learns the functional (linear) relationship between the belief state and the next-token probabilities, instead of memorizing.

M Complexification through coarse-graining: construction and dimensions

This appendix constructs an explicit two-agent feedback loop for which coarse-graining the observed interface by means of hiding either agent's emissions produces a measurable and predictable increase in the belief-state dimension an optimal predictor must represent. The increase in dimension happens because while the fully observed feedback loop factorizes, the coarse-grained process cannot make use of the same compression, as it does not fully observe all exchanged tokens between the two factors.

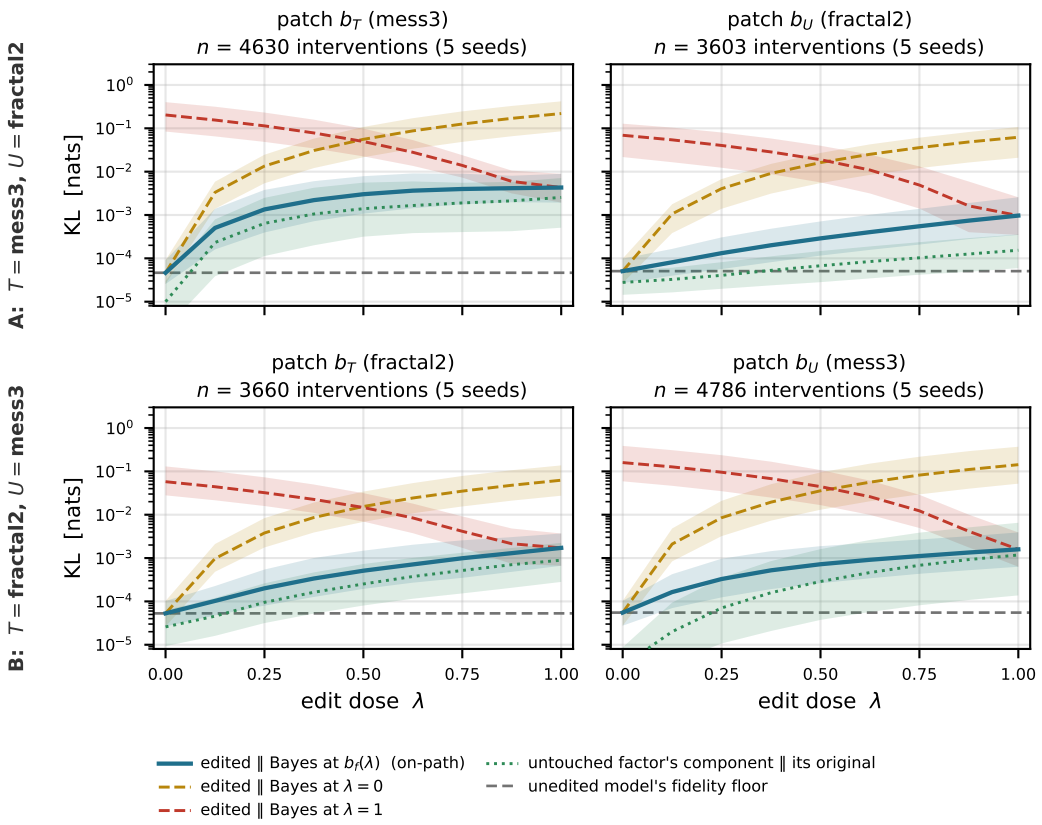


Figure 17: **Belief edits track the Bayes prediction continuously along the interpolation path.** KL divergence (median over interventions; bands: interquartile range) between the edited next-token distribution and three references, as a function of the edit dose λ: the Bayes prediction at the interpolated tuple $(b_f(\lambda), b_{-f})$ (on-path, solid), and the Bayes predictions at the two endpoints ($\lambda = 0$: original; $\lambda = 1$: full counterfactual; dashed). Rows: the two feedback setups; columns: the patched factor; n: evaluable interventions pooled over five seeds per setup. The dashed horizontal rule is the unedited model's own fidelity floor; the dotted curve is the untouched factor's component against its original prediction (leakage). At $\lambda = 0.5$ the nearest endpoint reference is 16–56$\times$ farther from the edited output than the interpolated target. Panel counts differ only through the identifiability gate (0.05 total variation between the endpoint references): the candidate pool is identical across panels, but the fractal2 belief distribution concentrates on a near-degenerate one-dimensional set, so swapping a fractal2 belief more often produces indistinguishable references.

To demonstrate this complexification, we analyzed a feedback loop between two transducers implementing a noisy, input-dependent permutation of their five internal states, with a noisy mapping to three emission tokens for each transition. Fig. 18 illustrates the construction and the three observed interfaces. For a factor with states $\mathcal{S} = \{0, \ldots, 4\}$, input symbol $k \in \{0, 1, 2\}$ (the partner's current symbol) and output alphabet $\{0, 1, 2\}$:

$$T^{(k)} = (1 - \varepsilon)\, P_{\pi_k} + \tfrac{\varepsilon}{5}\, \mathbf{1}\mathbf{1}^\top, \qquad \Pr(v \mid s') = (1 - \delta)\, [\rho(s') = v] + \tfrac{\delta}{3}, \tag{65}$$

where P_{π_k} is the permutation matrix of π_k, the emission is read from the *arrival* state s', and $\rho : \mathcal{S} \to \{0, 1, 2\}$ is a fixed 5-to-3 partition map. The permutations π_k, the partition maps ρ and the noise levels ($\varepsilon \in [0.02, 0.1]$, $\delta \in [0, 0.15]$) of both agents are drawn independently for each replicate of Appendix M.2; Fig. 18a shows one kernel of the draw analyzed in Appendix M.1. The joint token is the pair (x, y) of both emissions (vocabulary 9); the two coarse-grainings observe only x or only y (vocabulary 3), marginalizing the other symbol while its dynamics continue to act.

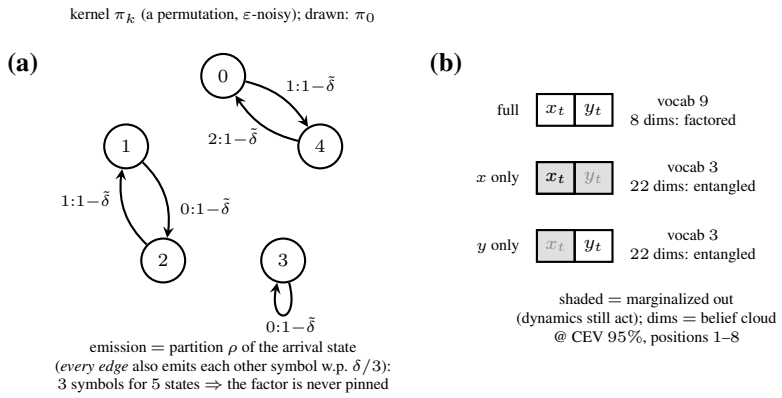


Figure 18: **The coarse-graining demonstrator process:** two five-state transducers of this family in the canonical feedback loop, each one's emitted symbol the other's next input symbol k (agent action-delayed). **(a)** Structure of one factor (edge labels read "emission : probability" as in Fig. 14, with $\tilde{\delta} \equiv 2\delta/3$). Each input symbol k selects a near-deterministic permutation kernel of the five states (drawn: $\pi_0 = (4, 2, 1, 3, 0)$, two transpositions and a fixed point), so different hidden-action histories imply visibly different dynamics; each transition emits only a three-way *partition* ρ of its arrival state (edges into the same state carry the same dominant token), so an observer never pins the state exactly and cannot invert the partner's symbol. **(b)** The three observed interfaces. Observing the pair (x_t, y_t) makes the joint belief factorize into two per-factor beliefs (8 dimensions); marginalizing either symbol destroys the factorization and forces an entangled belief over the 75-state joint latent (22 dimensions), even though the coarse observer receives strictly less information.

M.1 Theoretical and measured dimensions

The dimensions of the factored simplices can be computed directly: agent T has a dimension of $d_1 = 5$, while the action-delayed agent U has dimension $d_2 = 5 \cdot 3 = 15$ (mind $\times$ latched action). In contrast, the Kronecker product of the two simplices lives on a 74-dimensional simplex (75 states), revealing a large dimensional compression through factorization in this process:

$$\underbrace{\textstyle\sum_f (d_f - 1) = 18}_{\text{factored budget}} \ll \underbrace{\textstyle\prod_f d_f - 1 = 74}_{\text{joint simplex (hull bound)}}, \qquad \text{full-obs. product cloud} \;\leq\; 74. \tag{66}$$

Exact filtering on the 75-state joint machine (labeled transition operators extracted from the factored implementation by delta-probing; 1536 sequences) gives, within the model context window (positions 1–8), cumulative-explained-variance dimensions:

observation	representation	dims @80%	dims @95%
full (x, y)	factored concatenation $(b^{(1)}, b^{(2)})$	7	8
full (x, y)	joint product cloud (containment bound)	19	23
only x (agent hidden)	entangled joint belief	16	22
only y (env. hidden)	entangled joint belief	17	22

Both coarse-grainings force the observer from the 8-dimensional factored representation onto a 22-dimensional entangled cloud — a $2.75\times$ increase at the 95% threshold, obtained by *removing* observation symbols — while respecting containment in the fully-observed product cloud ($22 \leq 23$), confirming that the mechanism is the loss of factored compression rather than the creation of new belief geometry. (Pooled over longer horizons the coarse clouds saturate the hull, dims 19/23 at 80/95%.) The prediction for trained networks is the corresponding effective-rank inversion between models trained on the full (9-token) and coarse (3-token) streams of the same process.

M.2 Trained networks: the effective-rank inversion

We trained transformers, identical in architecture and optimization to all other runs (Table 2), on the three streams of five independent draws of this process family (replicates r0–r4; r0 is the draw analyzed in Appendix M.1, and each draw has its own permutation kernels, partition maps and noise levels (ε, δ) for both agents). For each draw we trained one model per interface — the full joint

Table 3: **Effective rank of the final-layer residual stream per replicate**, the smallest number of principal components whose CEV reaches 95% (resp. 99%), on the three observed interfaces of the coarse-graining demonstrator process (1024 sequences, BOS position excluded). Replicates are independent draws of the process family (r0 is the draw analyzed in Appendix M.1). For reference, the exact Bayes beliefs on the same sequences reach 95% at 8 dimensions under full observation in every draw and at 18–22 dimensions under either coarse-graining.

	full (x, y)		only x (U hidden)		only y (T hidden)	
replicate	95%	99%	95%	99%	95%	99%
r0	16	48	32	69	24	56
r1	21	56	42	77	27	57
r2	13	38	27	62	25	55
r3	10	29	12	38	23	54
r4	20	53	27	62	34	67
mean	16.0	44.8	28.0	61.6	26.6	57.8

token (vocabulary 9) and the two coarse-grainings (vocabulary 3) — with its own training seed, 15 models in all. All 15 converge to their own Bayes-optimal floors (final loss within $\pm 4.4 \times 10^{-3}$ nats of the floor, i.e. at the Monte-Carlo precision of the floor estimates), so the comparison is between models that have solved their respective prediction tasks. The exact belief dimensions are essentially draw-independent (8 under full observation in every draw; 20–22 resp. 18–22 at 95% CEV under the two coarse-grainings), whereas the trained models' ranks vary considerably between draws, so the within-draw comparison of the three interfaces is the relevant statistic.

Figure 4e and Table 3 confirm the predicted inversion. At the 95% threshold the final-layer residual stream has a mean effective rank of 16 (range 10–21) for the fully observing models but 28 (range 12–42) and 27 (range 23–34) for the models observing only x resp. only y (at 99%: 45 against 62 and 58). The absolute ranks vary considerably between draws, but the within-draw difference does not change sign: in all ten (draw, coarse interface) pairs the coarse model's rank exceeds that of the fully observing model of the same draw, by 2–21 components at 95% and 1–25 at 99%. The networks given strictly less information thus carry the higher-dimensional representation, as they cannot use the factored compression.

Three limitations should be stated. First, the inversion is visible at the 95% and 99% thresholds but not at 80% (mean ranks 5.8, 6.4 and 6.2; the within-draw difference changes sign across draws). Second, the effect is weakest in draw r3, where the x-only model exceeds the full-observation model by only two components at 95%. Third, model ranks exceed the corresponding belief ranks everywhere (e.g. 16 vs 8 under full observation), so the raw effective rank is a loose upper proxy for the dimension of the belief representation itself; isolating the belief-predictive subspace before measuring its rank would sharpen the comparison.

M.3 Computational Mechanics

Recall that if we have two interfaces

$$I_T[y_{0:L}|x_{0:L}] = \langle 1_R| \prod_{t=0}^{L-1} \hat{T}^{(y_t|x_t)} |\rho_R\rangle, \tag{67}$$

and

$$I_U[x_{0:L}|y_{0:L}] = \sum_{x_L} p(x_0) \langle 1_S| \prod_{t=0}^{L-1} \hat{U}^{(x_{t+1}|y_t)} |\rho_S\rangle, \tag{68}$$

the joint process that arises from a perception-action loop between them is

$$\Pr(X_{0:L} = x_{0:L}, Y_{0:L} = y_{0:L}) \tag{69}$$

$$= I_T[y_{0:L}|x_{0:L}] I_U[x_{0:L}|y_{0:L}] \tag{70}$$

$$= \sum_{x_L} p(x_0) \langle 1_{S\otimes R}| \prod_{t=0}^{L-1} \left(\hat{U}^{(x_{t+1}|y_t)} \otimes \hat{T}^{(y_t|x_t)} \right) |\rho_{S\otimes R}\rangle, \tag{71}$$

where we have used the notation $|v_{R\otimes S}\rangle \equiv |v_R\rangle \otimes |v_S\rangle$ (Fiderer et al., 2025).

What is the relationship between the ϵ-machine of the joint process and the ϵ-transducers of each individual process? All of these operate on histories of actions x and perceptions y. It's worth noting that while the interface defines the conditional probability whenever it is used in a feed-forward context, it is not necessarily true if there is feedback (information flows from the transducer's outputs to its inputs) (Boyd et al., 2025). For this reason, when we specify probabilities, we will indicate a subscript: $\Pr_J$ to indicate the joint feedback process, $\Pr_T$ to indicate feed-forward transduction from x to y, and $\Pr_U$ to indicate feed-forward transduction from y to x. Concretely, we have three separate equivalence conditions

1. *Joint Process:* Two histories $x_{0:L}y_{0:L}$ and $x'_{0:L}y'_{0:L}$ are equivalent iff they produce equivalent future distributions $X_{L:\infty}Y_{L:\infty}$:

$$\epsilon_J(x_{0:L}y_{0:L}) = \epsilon_J(x'_{0:L}y'_{0:L}) \tag{72}$$
$$\Leftrightarrow \tag{73}$$
$$\Pr_J(X_{L:\infty}, Y_{L:\infty}|X_{0:L} = x_{0:L}, Y_{0:L} = y_{0:L}) = \Pr_J(X_{L:\infty}, Y_{L:\infty}|X_{0:L} = x'_{0:L}, Y_{0:L} = y'_{0:L}) \tag{74}$$
$$\frac{I_T[y_{0:L}y_{L:\infty}|x_{0:L}x_{L:\infty}]I_U[x_{0:L}x_{L:\infty}|y_{0:L}y_{L:\infty}]}{I_T[y_{0:L}|x_{0:L}]I_U[x_{0:L}|y_{0:L}]} = \frac{I_T[y'_{0:L}y_{L:\infty}|x'_{0:L}x_{L:\infty}]I_U[x'_{0:L}x_{L:\infty}|y'_{0:L}y_{L:\infty}]}{I_T[y'_{0:L}|x'_{0:L}]I_U[x'_{0:L}|y'_{0:L}]}. \tag{75}$$

2. *X to Y:* Two histories $x_{0:L}y_{0:L}$ and $x'_{0:L}y'_{0:L}$ are equivalent iff they produce equivalent conditional future distributions $Y_{L:\infty}|X_{L:\infty}$:

$$\epsilon_T(x_{0:L}y_{0:L}) = \epsilon_T(x'_{0:L}y'_{0:L}) \tag{76}$$
$$\Leftrightarrow \tag{77}$$
$$\Pr_T(Y_{L:\infty}|X_{L:\infty}, X_{0:L} = x_{0:L}, Y_{0:L} = y_{0:L}) = \Pr_T(Y_{L:\infty}|X_{L:\infty}, X_{0:L} = x'_{0:L}, Y_{0:L} = y'_{0:L}) \tag{78}$$
$$\frac{I_T[y_{0:L}y_{L:\infty}|x_{0:L}x_{L:\infty}]}{I_T[y_{0:L}|x_{0:L}]} = \frac{I_T[y'_{0:L}y_{L:\infty}|x'_{0:L}x_{L:\infty}]}{I_T[y'_{0:L}|x'_{0:L}]}. \tag{79}$$

3. *Y to X:* Unlike the other two cases, the histories of relevance to the transducer U are time-shifted. Two histories $x_{0:L+1}y_{0:L}$ and $x'_{0:L+1}y'_{0:L}$ are equivalent iff they produce equivalent conditional future distributions $Y_{L:\infty}|X_{L:\infty}$:

$$\epsilon_U(x_{0:L+1}y_{0:L}) = \epsilon_U(x'_{0:L+1}y'_{0:L}) \tag{80}$$
$$\Leftrightarrow \tag{81}$$
$$\Pr_U(X_{L+1:\infty}|Y_{L:\infty}, X_{0:L+1} = x_{0:L+1}, Y_{0:L} = y_{0:L}) = \Pr_U(X_{L+1:\infty}|Y_{L:\infty}, X_{0:L+1} = x'_{0:L+1}, Y_{0:L} = y'_{0:L}) \tag{82}$$
$$\frac{I_U[x_{0:L+1}x_{L+1:\infty}|y_{0:L}y_{L:\infty}]}{I_U[x_{0:L+1}|y_{0:L}]} = \frac{I_U[x'_{0:L+1}x_{L+1:\infty}|y'_{0:L}y_{L:\infty}]}{I_U[x'_{0:L+1}|y'_{0:L}]}. \tag{83}$$

This breakdown illustrates why the belief states of the joint process are not simply the tensor product of the belief states of T and U. If it were, then the equivalence map ϵ_J would identify histories as the same if both ϵ_T and ϵ_U do. However, we see a subtlety here due to time-indexing. Let us assume that the joint ϵ-machine does indeed store the ϵ-transducer states of T. Then two histories being equivalent would imply

$$\frac{I_T[y_{0:L}y_{L:\infty}|x_{0:L}x_{L:\infty}]}{I_T[y_{0:L}|x_{0:L}]} = \frac{I_T[y'_{0:L}y_{L:\infty}|x'_{0:L}x_{L:\infty}]}{I_T[y'_{0:L}|x'_{0:L}]}. \tag{84}$$

and therefore

$$\frac{I_U[x_{0:L}x_{L:\infty}|y_{0:L}y_{L:\infty}]}{I_U[x_{0:L}|y_{0:L}]} = \frac{I_U[x'_{0:L}x_{L:\infty}|y'_{0:L}y_{L:\infty}]}{I_U[x'_{0:L}|y'_{0:L}]}, \tag{85}$$

which is the causal equivalence relation for the interface that produces $X_{0:t}$ from $Y_{0:t}$. This is importantly different from the ϵ-transducer that would correspond to the belief states of U, which would produce the process $X_{1:t+1}$ from $Y_{0:t}$. This is essentially the U transducer with a time-lag in the action output, which may require more memory. Thus, it is sufficient to model the belief-states of T and time-lagged U separately to fully capture the minimal predictive states of the feedback process.

These relations give us access to the statistical complexity, which is the entropy of the causal states. Let the causal states be denoted:

1. $S_t^J = \epsilon_J(\overleftarrow{x}_t, \overleftarrow{y}_t)$ for the joint process.
2. $S_t^T = \epsilon_T(\overleftarrow{x}_t, \overleftarrow{y}_t)$ for the interface that agent T implements $I_T[y_{0:L}|x_{0:L}]$.
3. $S_t^{\mathcal{U}} = \epsilon_{\mathcal{U}}(\overleftarrow{x}_t, \overleftarrow{y}_t)$ for the interface that action-delayed agent U implements $I_{\mathcal{U}}[x_{0:L}|y_{0:L}]$.

The exploration in the previous section establishes that the joint causal states are a function of the causal states of the ϵ-transducers

$$S_t^J = f(S_t^T, S_t^{\mathcal{U}}). \tag{86}$$

Therefore, the complexity of the joint process is bounded above by the uncertainty in causal states of both interfaces:

$$C_\mu^J \equiv H[S_t^J] \leq H[S_t^T, S_t^{\mathcal{U}}]. \tag{87}$$