Parameter Learning in Active Inference via Variational Message Passing

Benedikt Waldeck[1 [0009−0001−5225−9933]]

University of Modena and Reggio Emilia, Italy
benedikt.waldeck@unimore.it

Abstract. Active inference casts perception, learning and planning as variational inference, where the Expected Free Energy (EFE) defines the planning objective. Recent work showed that EFE minimization can be written as Variational Free Energy (VFE) minimization on a generative model augmented with epistemic priors, and that the resulting entropy-corrected objective is minimized by sum-product message passing (SP-MP) on modified factor kernels. These formulations treat model uncertainty as *inference* over a finite set of known models. We extend them to planning during *learning*, where beliefs over the observation and transition parameters θ_A, θ_B are Dirichlet priors. Closed-form Dirichlet updates are usually obtained by a structured mean-field factorization between parameters and states, but under exactly that factorization the information-gain corrections vanish. We resolve this by giving every factor its own parameter prior, frozen during a rollout. Conjugacy then makes each factor belief a discrete belief times a one-step Bayes posterior, so the observation-dependent parameter posterior that mean-field destroys is retained. We show that evaluating the epistemic priors under the resulting Bethe joint recovers the information gains, and that stationarity of this single closed-form objective yields SP-MP whose factor kernels combine predictive likelihoods with multiplicative information-gain weights. Experiments on two grid POMDPs, in which only a planner that learns the model finds the safe route, show that the method plans to efficiently learn θ_A and θ_B and reaches the goal on every episode, whereas SP-MP planners with the same inference update plateau or take the risky route.

Keywords: Active Inference · Parameter Learning · Planning as Inference · Factor Graphs · Variational Inference

1 Introduction

Active inference casts perception, planning and learning as variational inference on a single generative model [3]. Its planning objective, the Expected Free Energy (EFE), decomposes into risk, ambiguity and novelty and thereby balances goal pursuit against epistemic exploration, but evaluating it by enumerating policies scales exponentially with the horizon [5]. De Vries et al. [11] showed that EFE

minimization can be reformulated as Variational Free Energy (VFE) minimization on a generative model augmented with *epistemic priors*. Nuijten et al. [14] showed that the augmented VFE equals the predictive VFE plus explicit *entropy corrections*, including the planning correction of Lázaro-Gredilla et al. [10], and derived a message passing scheme in which each correction becomes a *channel reparameterization* of a factor kernel, so that EFE-based planning runs as sum-product message passing (SP-MP) on the original factor graph.

These formulations treat the model parameter θ as a single enumerable latent variable: they plan under uncertainty over a finite set of known models, not under uncertainty about the models themselves. Planning during *Learning* the observation model θ_A and the transition model θ_B conjugately, the setting of an agent that knows neither its sensors nor its actuators, is left open and poses a structural problem. Closed-form Dirichlet updates require a mean-field factorization between the parameter beliefs and the state–observation blocks, but under exactly that factorization every parameter-posterior contrast collapses, $q(\theta \mid \cdot) = q(\theta)$, and the novelty term that motivates parameter learning vanishes from the objective.

This paper resolves this problem within a single variational objective. We split θ into an observation block θ_A, a transition block θ_B and an enumerable context θ_c, and carry the entropy-corrected formulation through to conjugate learning of either block. Our contributions are:

- We show that the channels of the entropy corrections need not be free variational parameters. Substituting them at their closed-form optima, the conditionals of q, and giving each factor a step-wise parameter copy frozen at the current belief, keeps every sum-product factor belief conjugate and preserves the local parameter–outcome joints that mean-field destroys. Under the resulting Bethe joint the objective is a single scalar over the discrete coordinates, in which the parameter-side terms cancel into predictive likelihoods while the information-gain corrections survive in full.
- We derive the resulting message passing scheme. Four epistemic priors recover the EFE with separate novelty terms for θ_A and θ_B and a salience term for θ_c. Stationarity yields belief propagation on the discrete graph with self-consistent channel factors and per-entry *information-gain weights* $e^{D_{\mathrm{KL}}} \geq 1$ that emerge from the variation rather than being imposed; the fixed-parameter scheme of [13, 14] is recovered in the point-mass limit.
- We validate the method on two grid POMDPs in which only a planner that explores the model finds the safe route. The EFE-planner learns θ_A and θ_B and reaches the goal on every episode, whereas SP-MP baselines with the same filter and count update plateau or gamble; enumerated EFE reproduces its behaviour, and ablations isolate the novelty terms as the drive.

Section 2 reviews the EFE and the Bethe free energy. Section 3 discusses related work. Section 4 derives the objective and the message passing scheme. Section 5 reports the experiments and Section 6 concludes.

2 Background

2.1 Active Inference and the Expected Free Energy

We consider an agent that maintains a generative model predicting future observations, states and the consequences of actions over a planning horizon T. For a discrete partially observable Markov decision process (POMDP) with hidden states $\boldsymbol{x}$, observations $\boldsymbol{y}$, actions $\boldsymbol{u}$ and model parameters θ, we write this as a rollout model $p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) = p(\theta)\, p(x_0) \prod_{t=1}^{T} p(u_t)\, p(y_t \mid x_t, \theta)\, p(x_t \mid x_{t-1}, u_t, \theta)$. To encode goals, the model is augmented with *preference priors* $\hat{p}(x_t)$, $\hat{p}(y_t)$ over desired states and observations, which can be understood as exponentiated rewards $\hat{p} \propto \exp(R)$ [3].

Active inference selects actions by minimizing the Expected Free Energy (EFE), which scores a policy by the cost of its predicted trajectory and decomposes into three interpretable terms [3]. Per time step,

$$G(u_t) = \underbrace{D_{\mathrm{KL}}\big[q(x_t \mid u_t) \,\|\, \hat{p}(x_t)\big]}_{\text{risk (states)}} + \underbrace{\mathbb{E}_{q(x_t \mid u_t)}\big[\mathbb{H}[p(y_t \mid x_t)]\big]}_{\text{ambiguity}} - \underbrace{\mathbb{E}_{p(y_t \mid x_t)\, q(x_t \mid u_t)}\big[D_{\mathrm{KL}}[q(\theta \mid y_t, x_t) \,\|\, q(\theta)]\big]}_{\text{novelty}}. \tag{1}$$

Risk penalizes deviation of the predicted states from the preferences, ambiguity penalizes states whose observations are uncertain, and novelty rewards states and observations that are informative about θ. In discrete active inference, policies (sequences of actions) are typically scored with (1) and selected in proportion to $\exp(-G)$ [5].

2.2 Variational Free Energy and Message Passing on Factor Graphs

Given a generative model p, variational inference approximates its posterior by minimizing the Variational Free Energy (VFE) over a family of tractable distributions q:

$$F[q] = D_{\mathrm{KL}}\big[q(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) \,\|\, p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)\big]. \tag{2}$$

The generative model factorizes into local terms and can be represented as a Forney-style factor graph (FFG) [18], in which nodes $a \in \mathcal{V}$ represent factors f_a, edges $i \in \mathcal{E}$ represent variables, and d_i is the degree of edge i. The *Bethe approximation* constrains q to respect this structure: each node maintains a belief $q_a(\boldsymbol{s}_a)$ over its adjacent variables and each edge a singleton belief $q_i(s_i)$, with

$$q(s) = \frac{\prod_a q_a(\boldsymbol{s}_a)}{\prod_i q_i(s_i)^{d_i - 1}}, \qquad \mathbb{H}_{\mathrm{Bethe}}[q] = \sum_a \mathbb{H}[q_a] - \sum_i (d_i - 1)\, \mathbb{H}[q_i], \tag{3}$$

so that the joint entropy $\mathbb{H}[q]$ is approximated by $\mathbb{H}_{\mathrm{Bethe}}[q]$ and the VFE reduces to the *Bethe free energy*

$$F_{\mathrm{Bethe}}[q] = \sum_{a \in \mathcal{V}} D_{\mathrm{KL}}\big[q_a(\boldsymbol{s}_a) \,\|\, f_a(\boldsymbol{s}_a)\big] + \sum_{i \in \mathcal{E}} (d_i - 1)\, \mathbb{H}[q_i(s_i)]. \tag{4}$$

The beliefs must be normalized and locally consistent, so minimizing (4) is a constrained problem whose Lagrangian stationarity conditions yield message passing: sum-product message passing (belief propagation) under the consistency constraints alone, and Variational Message Passing (VMP) when node beliefs are further mean-field factorized [18].

3 Related Work

Planning-as-inference and active inference. Active inference inherits the planning-as-inference framework [1, 17, 9] and extends it by treating information gain as an intrinsic objective, so that the EFE combines pragmatic and epistemic value [6, 3]. Existing EFE planners employ specialized procedures: policy tree search [4], branching-time recursion [2], dynamic programming [16], or message passing on a *generalized* free energy [15] with a mutual-information term over future graph segments [8]; none minimizes a single variational objective over the whole graph.

EFE minimization as VFE minimization. Control-as-inference [7, 9] augments p with preference priors $\hat{p}(x_t)$ and, where used, $\hat{p}(y_t)$, giving a preference-biased model $\hat{p}$ and the VFE $F_{\hat{p}}[q] = \mathbb{E}_q[\log q/\hat{p}]$. De Vries et al. [11] (Proposition 1) showed that further augmenting $\hat{p}$ with *epistemic priors* $\tilde{p}$ makes this VFE equal to $\mathbb{E}_{q(\boldsymbol{u})}[G(\boldsymbol{u})]$ up to a complexity term. Nuijten et al. [14] showed that the augmented VFE can equivalently be written as $F_{\hat{p}}$ plus explicit *entropy corrections*, one conditional-entropy term per prior: $\tilde{p}(u_t)$ contributes $-\mathbb{H}[q(x_t \mid x_{t-1}, u_t)]$ and the observation-side priors contribute $+\mathbb{H}[q(y_t \mid x_t, \theta)]$ terms. Proper planning over policies requires one further *planning correction* $+\mathbb{H}[q(u_t \mid x_{t-1})]$, introduced by Lázaro-Gredilla et al. [10] as the correction that turns marginal inference into policy optimization [14].

Message passing realizations. The conditional entropies above are ratios of region beliefs and therefore not independent coordinates of the Bethe free energy. Nuijten et al. [13] attached the epistemic priors to a parameter-free model as extra factor nodes and resolved their circular dependence on q by alternating posterior and prior updates. Nuijten et al. [12] introduced region-extended coordinates and a single observation-side channel for $\mathbb{H}[q(y_t \mid x_t, \theta)]$, without the planning correction. Nuijten et al. [14] finally absorbed every correction, including the planning correction, into modified factor kernels through channel reparameterization, recovering belief propagation, cross-entropy planning and full EFE-based planning as one family. The key identity of that construction is the variational characterization of conditional entropy (Gibbs' inequality),

$$\mathbb{H}[q(y \mid x)] = \min_{r} \mathbb{E}_{q(y,x)}\big[-\log r(y \mid x)\big], \tag{5}$$

with equality at the normalized *channel* $r(y \mid x) = q(y \mid x)$. Treating each channel as a free variational parameter turns a single factor into a modified *kernel*, and the corrected Bethe objective is minimized by sum-product updates on those kernels, jointly over beliefs and channels.

This paper. Every realization above assumes a single, enumerable θ, and carries the channels of (5) as free variational parameters. We instead let θ carry Dirichlet beliefs over the model parameters themselves, and *eliminate* the channels at their closed-form optima, where conjugate factor beliefs supply the parameter–datum coupling that the channels would otherwise have to represent.

4 Methodology

4.1 Generative Model and Planning Problem

We consider a discrete-time POMDP with hidden states x_t, actions u_t, observations y_t and an episode-level parameter $\theta = (\theta_A, \theta_B, \theta_c)$ in three blocks,

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta) &= p(\theta_A)\, p(\theta_B)\, p(\theta_c)\, p(x_0) \prod_{t=1}^{T} p(u_t)\, f_{\mathrm{dyn},t}\, f_{\mathrm{obs},t}, \\ f_{\mathrm{dyn},t} &= p(x_t \mid x_{t-1}, u_t, \theta_B, \theta_c), \qquad f_{\mathrm{obs},t} = p(y_t \mid x_t, \theta_A, \theta_c). \end{aligned} \tag{6}$$

The context $\theta_c \in \{1, \dots, K\}$ is Categorical and selects which of K observation and transition matrices applies; θ_A, θ_B collect those matrices, $p(y_t \mid x_t, \theta_A, \theta_c) = \theta_A(y_t \mid x_t, \theta_c)$ and $p(x_t \mid x_{t-1}, u_t, \theta_B, \theta_c) = \theta_B(x_t \mid x_{t-1}, u_t, \theta_c)$, with one Dirichlet prior per column, $p(\theta_A) = \prod_{x,\theta_c} \mathrm{Dir}(\alpha_A(\cdot \mid x, \theta_c))$ and $p(\theta_B) = \prod_{x,u,\theta_c} \mathrm{Dir}(\alpha_B(\cdot \mid x, u, \theta_c))$. Unknown θ_A, θ_B means unlearned matrices; unknown θ_c means known matrices but an unknown context, the enumerable model variable of [14]. One draw of θ is shared by all T factors. Goals enter as preference priors, giving the preference-biased model $\hat{p} = p \prod_t \hat{p}_x(x_t)\, \hat{p}_y(y_t)$ (Figure 1) with VFE $F_{\hat{p}}[q] = \mathbb{E}_q[\log q/\hat{p}]$ ($\hat{p}$ unnormalized, so a divergence up to a constant).

The agent scores an action by the per-step EFE with a θ-free dynamics numerator,

$$\begin{aligned} G(u_t) = \mathbb{E}_{q(\cdot \mid u_t)}\Big[&\log \frac{q(x_t \mid x_{t-1}, u_t)}{\hat{p}_x(x_t)\, \hat{p}_y(y_t)\, p(y_t \mid x_t, \theta_A, \theta_c)} \\ &+ \log \frac{q(\theta_A, \theta_c \mid x_t)}{q(\theta_A, \theta_c \mid y_t, x_t)} + \log \frac{q(\theta_B \mid x_{t-1}, \theta_c)}{q(\theta_B \mid x_t, x_{t-1}, u_t, \theta_c)} \Big], \end{aligned} \tag{7}$$

where $q(\cdot \mid u_t)$ abbreviates $q(y_t, x_t, x_{t-1}, \theta \mid u_t)$. The first term is risk plus ambiguity $\mathbb{H}_q[y_t \mid x_t, \theta_A, \theta_c]$. The remaining terms contrast each block's *cavity*, the belief before the imagined datum, with its *posterior*, and are information gains. By the chain rule they split into the *salience* $\mathcal{I}_{c,t} = I(\theta_c; y_t \mid x_t)$, which rewards resolving the context, and the *novelties* $\mathcal{I}_{A,t} = I(\theta_A; y_t \mid x_t, \theta_c)$ and $\mathcal{I}_{B,t} = I(\theta_B; x_t, u_t \mid x_{t-1}, \theta_c)$, which reward learning the matrices. The planning problem is to find the policy $q(u_t \mid x_{t-1})$ that minimizes $\sum_t \mathbb{E}_{q(u_t)}[G(u_t)]$.

Two obstacles separate (7) from a message-passing algorithm. First, G contains the agent's own conditionals inside the logarithm, so it is not a VFE against any fixed model. Second, the novelties are mutual informations of *local joints* such as $q(y_t, x_t, \theta_c, \theta_A)$; a family that factorizes the Dirichlet block away from

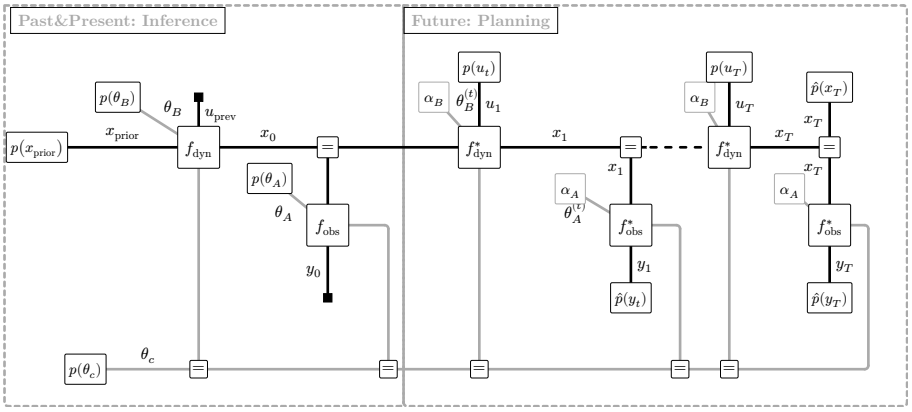


Fig. 1. Frozen-parameter factor graph (Section 4.3). Squares: factors; edges: variables; =: equality nodes; filled squares: clamped data; $\hat{p}$: preferences. *Left:* the real slice carries the shared θ_A, θ_B and turns $p(x_{\text{prior}}), u_{\text{prev}}, y_0$ into the planning prior $p(x_0)$. *Right:* over the horizon each factor f^* (kernels (17)–(18)) owns a copy $\theta_A^{(t)}, \theta_B^{(t)}$ with its own degree-two prior node (open, gray) frozen at $\text{Dir}(\alpha_A)$ resp. $\text{Dir}(\alpha_B)$; only θ_c is shared, through equality nodes.

the imagined datum, $q(y_t, x_t, \theta_c)\, q(\theta_A)$, has $\mathcal{I}_{A,t} \equiv 0$, yet this is exactly the factorization that structured mean-field uses to obtain Dirichlet conjugacy. Section 4.2 removes the first, Section 4.3 the second.

4.2 From Expected to Variational Free Energy

Following [11, 14], we keep the objective and change the model: $\hat{p}$ is augmented with *epistemic priors*, one for every factor of (7) that is not already part of $\hat{p}$,

$$\tilde{p}(u_t) \propto \exp\big(-\mathbb{E}_{q(x_t, x_{t-1} \mid u_t)} \log q(x_t \mid x_{t-1}, u_t)\big), \tag{8}$$

$$\tilde{p}(x_t) \propto \exp\big(-\mathbb{E}_{q(\theta_A, \theta_c \mid x_t)}\, h[q(y_t \mid x_t, \theta_A, \theta_c)]\big), \tag{9}$$

$$\tilde{p}(y_t, x_t) \propto \exp D_{\text{KL}}[q(\theta_A, \theta_c \mid y_t, x_t) \,\|\, q(\theta_A, \theta_c \mid x_t)], \tag{10}$$

$$\tilde{p}(x_t, x_{t-1}, u_t, \theta_c) \propto \exp D_{\text{KL}}[q(\theta_B \mid x_t, x_{t-1}, u_t, \theta_c) \,\|\, q(\theta_B \mid x_{t-1}, \theta_c)], \tag{11}$$

where $h[q(\cdot \mid x)]$ denotes the entropy of a conditional as a function of x and $\mathbb{H}_q[\,\cdot \mid x] = \mathbb{E}_{q(x)} h[q(\cdot \mid x)]$ the conditional entropy. Each prior is the exponential of a local functional of q, so $-\mathbb{E}_q[\log \tilde{p}]$ is closed-form.

Corollary 1 (Epistemic priors for three parameter blocks; cf. [11], Prop. 1). *Let $\tilde{p}$ be $\hat{p}$ augmented with the priors* (8)–(11) *for $t = 1, \ldots, T$. Then $F_{\tilde{p}}[q] = \sum_t \mathbb{E}_{q(u_t)}[G(u_t)] + B[q]$, where $B[q] = D_{\text{KL}}[q \,\|\, p] \geq 0$ is the complexity of q relative to* (6).

Proof. See Appendix A.1.

The prior (11) is the transition-novelty prior, new relative to [14]. Proper planning over policies further requires the planning correction $+\mathbb{H}_q[u_t \mid x_{t-1}]$ [10,

14], which cancels the reward that the VFE pays for action entropy. Evaluating $-\mathbb{E}_q[\log \tilde{p}]$ for each prior (Appendix A.2) writes the EFE-based planning objective as the base VFE plus a per-step *entropy correction*, $F_T[q] = F_{\hat{p}}[q] + \sum_{t=1}^{T} \Delta F_t[q]$ (the subscript is the horizon) with

$$\Delta F_t[q] = \mathbb{H}_q[y_t \mid x_t] - 2\big(\mathcal{I}_{A,t} + \mathcal{I}_{c,t}\big) - \mathbb{H}_q[x_t \mid x_{t-1}, u_t] - \mathcal{I}_{B,t} + \mathbb{H}_q[u_t \mid x_{t-1}]. \quad (12)$$

The factor two is no double count: the ambiguity prior (9) pays one unit of each observation-side gain through its θ-coupling and the novelty prior (10) a second, while the transition side has no ambiguity term (Appendix A.2). Message passing on $\tilde{p}$ now applies, *provided* that q retains the local joints that make the information gains in (12) nonzero.

4.3 Variational Family: Sum-Product Beliefs on the Frozen-Parameter Graph

Frozen-parameter graph. The agent plans with its current parameter posterior, so imagined observations in the rollout must not update it. On the graph of (6) this fails: the sum-product message from $f_{\text{obs},t}$ to a shared θ_A node, $\sum_{y,x,\theta_c} p(y \mid x, \theta_A, \theta_c)\, \mu_y \mu_x \mu_{\theta_c}$ with μ the messages into the factor, is a mixture of monomials rather than a Dirichlet, so conjugacy is lost and imagined observations mix across the horizon. Both problems are resolved by a step-wise copy $\theta_A^{(t)}$ per observation factor with its own prior node $p(\theta_A^{(t)}) := q_{\theta_A} = \text{Dir}(\theta_A^{(t)} \mid \alpha_A)$ carrying the current real-data belief, and likewise $\theta_B^{(t)}$ per $f_{\text{dyn},t}$ (Figure 1; $q_{\theta_A}, q_{\theta_B}$ denote these beliefs, as opposed to the copy marginals). Setting a prior node to a belief makes $\tilde{p}$ a *planning-time* model, rebuilt each environment step, so the objective is single per planning call. Each copy has degree two, so the message into its factor is exactly $\text{Dir}(\alpha_A)$ and the return message reaches no other factor; the price is that novelty does not decay along the horizon, which replanning bounds (Appendix B.4). The finite θ_c is *not* copied: it is conjugate to any factor, and sharing it lets an early cue inform a distant step.

Factor beliefs are conjugate in closed form. A Bethe factor belief is the factor times its incoming messages, $q_{\text{obs},t} \propto p(y_t \mid x_t, \theta_A^{(t)}, \theta_c)\, \mu_{y_t} \mu_{x_t} \mu_{\theta_c} \text{Dir}(\theta_A^{(t)} \mid \alpha_A)$. For each context the likelihood is a monomial in one column of θ_A, so conjugacy applies entrywise, $\theta_A(y \mid x, \theta_c)\, \text{Dir}(\theta_A \mid \alpha_A) = m_A(y \mid x, \theta_c)\, \text{Dir}(\theta_A \mid \alpha_A + e_{yx\theta_c})$, with $e_{yx\theta_c}$ one count on entry y of column (x, θ_c) and $m_A = \alpha_A / \alpha_{A,0}$ the posterior predictive likelihood (Appendix A.3). Hence, exactly,

$$q_{\text{obs},t}(y_t, x_t, \theta_c, \theta_A^{(t)}) = q_{yx,t}(y_t, x_t, \theta_c)\, \underbrace{\text{Dir}(\theta_A^{(t)} \mid \alpha_A + e_{y_t x_t \theta_c})}_{q_{\text{obs},t}(\theta_A \mid x_t, y_t, \theta_c)}, \quad (13)$$

with discrete marginal $q_{yx,t} := \int q_{\text{obs},t}\, d\theta_A \propto m_A\, \mu_{y_t} \mu_{x_t} \mu_{\theta_c}$, and $q_{\text{dyn},t} = q_{\text{trip},t}(x_t, x_{t-1}, u_t, \theta_c)\, \text{Dir}(\theta_B \mid \alpha_B + e_{x_t x_{t-1} u_t \theta_c})$ with the predictive likelihood m_B of α_B.

Coordinates and constraints. The free coordinates are the Bethe beliefs of the discrete graph over $\{y_{1:T}, x_{0:T}, u_{1:T}, \theta_c\}$: the block beliefs $q_{yx,t}$, $q_{\mathrm{trip},t}$ and the singletons $q(x_t)$, $q(y_t)$, $q(u_t)$, $q(\theta_c)$, subject to normalization and factor-to-singleton consistency, with $q_{\theta_A} = \mathrm{Dir}(\alpha_A)$ and $q_{\theta_B} = \mathrm{Dir}(\alpha_B)$ held fixed during planning. The θ_A- and θ_B-conditionals of the factor beliefs are not coordinates: any other choice raises the free energy of the region, so they are substituted out at the closed-form optima of (13) (Appendix A.3, Lemma 3). There are no channel constraints either: the channels of (5) are not coordinates but definitions, $r = q(\cdot \mid \cdot)$.

Cavities and per-entry information gains. Novelty contrasts the posterior $q_{\mathrm{obs},t}(\theta_A \mid x_t, y_t, \theta_c)$ with the *cavity* $q_{\mathrm{obs},t}(\theta_A \mid x_t, \theta_c)$, the same belief before the outcome is known,

$$q_{\mathrm{obs},t}(\theta_A \mid x_t, \theta_c) = \sum_y q_{yx,t}(y \mid x_t, \theta_c)\,\mathrm{Dir}(\alpha_A + e_{yx_t\theta_c}), \tag{14}$$

a finite Dirichlet mixture, the only non-conjugate object in the construction. We project it onto $\mathrm{Dir}(\alpha_A + \bar{n}_{x_t\theta_c})$ with the expected fractional counts $\bar{n}_{x_t\theta_c}(y) := q_{yx,t}(y \mid x_t, \theta_c)$; it matches the mixture mean, is exact for one-hot $q_{yx,t}(\cdot \mid x_t, \theta_c)$, and is overconfident by $O(1/\alpha_{A,0})$ (Appendix A.4). The per-entry gains are then closed-form Dirichlet–Dirichlet divergences,

$$\Delta_{A,t}(x_t, y_t, \theta_c) = D_{\mathrm{KL}}\big[q_{\mathrm{obs},t}(\theta_A \mid x_t, y_t, \theta_c) \,\big\|\, q_{\mathrm{obs},t}(\theta_A \mid x_t, \theta_c)\big] \tag{15}$$

with $\mathcal{I}_{A,t} = \mathbb{E}_{q_{yx,t}}[\Delta_{A,t}]$, and symmetrically $\Delta_{B,t}(x_t, x_{t-1}, u_t, \theta_c)$ with cavity $q_{\mathrm{dyn},t}(\theta_B \mid x_{t-1}, \theta_c)$ and $\mathcal{I}_{B,t} = \mathbb{E}_{q_{\mathrm{trip},t}}[\Delta_{B,t}]$. Salience needs neither: θ_c is a stored axis, so $\mathcal{I}_{c,t} = I_{q_{yx,t}}(y_t; \theta_c \mid x_t)$ is a mutual information of the finite array.

4.4 The Planning Objective in Closed Form

Combining the factor beliefs along their separators gives the Bethe joint of the frozen graph, $q = \prod_t q_{\mathrm{obs},t}\, q_{\mathrm{dyn},t} \big/ \big(q(\theta_c)^{2T-1} \prod_{t<T} q(x_t)^2\, q(x_T)\big)$, with exponents from the variable degrees (Appendix A.5). Its θ-free conditionals are the block conditionals and its θ_A, θ_B-conditionals the posteriors and cavities of (13)–(14), so every object in (8)–(11) is well-defined under q.

Proposition 1 (Closed-form planning objective). *Under the Bethe joint q: (i) in the base free energy the parameter entropy of each factor belief cancels per entry against its parameter energy, leaving the Bethe free energy of the discrete graph with predictive energies $-\log m_A$, $-\log m_B$ (eq. (29)); (ii) the corrections (12) hold unchanged, with $\mathcal{I}_{A,t}, \mathcal{I}_{B,t}$ as in (15) and $\mathcal{I}_{c,t} = I_{q_{yx,t}}(y_t; \theta_c \mid x_t)$; (iii) hence the planning objective is the closed-form functional*

$$\begin{aligned} F_T[q] = {} & \textstyle\sum_{t=1}^{T} \Big(D_{\mathrm{KL}}\big[q_{yx,t} \,\big\|\, m_A\, \hat{p}_y\big] + D_{\mathrm{KL}}\big[q_{\mathrm{trip},t} \,\big\|\, m_B\, p(u_t)\, \hat{p}_x\big] \Big) + \mathcal{C}[q] \\ & + \sum_{t=1}^{T} \Big(\mathbb{H}_{q_{yx,t}}[y_t \mid x_t] - 2\,\mathbb{E}_{q_{yx,t}}[\Delta_{A,t}] - 2\, I_{q_{yx,t}}(y_t; \theta_c \mid x_t) \\ & \qquad - \mathbb{H}_{q_{\mathrm{trip},t}}[x_t \mid x_{t-1}, u_t] - \mathbb{E}_{q_{\mathrm{trip},t}}[\Delta_{B,t}] + \mathbb{H}_{q_{\mathrm{trip},t}}[u_t \mid x_{t-1}] \Big) \end{aligned} \tag{16}$$

Algorithm 1 Conjugate planning and learning. Inputs: counts α_A, α_B, beliefs $q(\theta_c), q(x_0)$, horizon T, damping η, tolerance ϵ, sweep budget N.

repeat ▷ one environment step
  m_A, m_B and one-count posteriors $\mathrm{Dir}(\alpha + e)$ from α_A, α_B ▷ once per call
  $r_t \leftarrow$ uniform, $\bar{n} \leftarrow$ prior predictive, $\mu_{\theta_c \to \cdot} \leftarrow q(\theta_c)$, $\mu_{x_0} \leftarrow q(x_0)$
  for $k = 1, \ldots, N$ **do**
    kernels $\bar{g}_{\mathrm{obs},t}, \bar{g}_{\mathrm{dyn},t}$; forward–backward pass ▷ (17)–(18)
    refresh $r_t \leftarrow q(\cdot \mid \cdot)$, $\bar{n}$ and the weights $\Delta_{A,t}, \Delta_{B,t}$, damped by η
    evaluate $F_T^{(k)}$; **break if** $|F_T^{(k)} - F_T^{(k-1)}| < \epsilon$ ▷ (16)
  end for
  execute $u_1 \in \arg\max q(u_1)$; observe y_1; filter $q(x_1)$; $q(\theta_c)$ by exact Bayes
  $\alpha_A \mathrel{+}= \mathbb{E}[e_{y_1 x_1 \theta_c}]$, $\alpha_B \mathrel{+}= \mathbb{E}[e_{x_1 x_0 u_1 \theta_c}]$ under $q(x_1, x_0)\, q(\theta_c)$; $q(x_0) \leftarrow q(x_1)$
until episode ends

over the discrete coordinates alone, where $\mathcal{C}[q]$ collects the Bethe counting terms and the x_0, θ_c priors.

Proof. See Appendix A.5.

4.5 Stationarity Conditions and Message Passing

Since F_T has Bethe form over the discrete coordinates, its stationarity conditions under the normalization and consistency multipliers (Appendix B.1) yield sum-product message updates. Two facts make the variation local and term-by-term (Appendix B.2): entropies of conditionals of a normalized joint differentiate as if the logarithm were frozen (Lemma 4), and the dependence of the cavity on $q_{yx,t}$ contributes nothing, $\delta \mathcal{I}_{A,t} / \delta q_{yx,t}(y, x, \theta_c) = \Delta_{A,t}(x, y, \theta_c)$ (Lemma 5).

Proposition 2 (Modified kernels). *At stationarity $q^*_{yx,t} \propto \bar{g}_{\mathrm{obs},t}$ and $q^*_{\mathrm{trip},t} \propto \bar{g}_{\mathrm{dyn},t}$, up to the singleton multipliers fixed in Appendix B.2, with the kernels*

$$\bar{g}_{\mathrm{obs},t}(y_t, x_t, \theta_c) = m_A(y_t \mid x_t, \theta_c)\, \frac{r_t(y_t \mid x_t, \theta_c)^2}{r_t(y_t \mid x_t)}\, e^{2\,\Delta_{A,t}(x_t, y_t, \theta_c)}, \tag{17}$$

$$\bar{g}_{\mathrm{dyn},t}(x_t, x_{t-1}, u_t, \theta_c) = m_B(x_t \mid x_{t-1}, u_t, \theta_c)\, \frac{r_t(u_t \mid x_{t-1})}{r_t(x_t \mid x_{t-1}, u_t)}\, e^{\Delta_{B,t}}, \tag{18}$$

evaluated at the self-consistent conditionals $r_t(\cdot \mid \cdot) = q^*(\cdot \mid \cdot)$ of the block beliefs.

Proof. See Appendix B.2, which also fixes the multipliers from the singleton stationarity conditions.

The kernels (17)–(18) are the kernel pair (13a)–(13b) of [14] with the point likelihood replaced by the predictive likelihood and an *information-gain weight* $e^{2\Delta_A}$, $e^{\Delta_B} \geq 1$ appended; the weights are not imposed but emerge from the variation of the information gains.

Algorithm, learning and the fixed-parameter limit. With r_t lagged by one sweep, as for the channels of [14], Proposition 2 is a standard Bethe belief propagation problem on the discrete graph over $\{y_t, x_t, u_t, \theta_c\}$ with kernels (17)–(18). Algorithm 1 solves it with closed-form updates (Appendix B.3–B.4); the copies $\theta_A^{(t)}, \theta_B^{(t)}$ exchange no messages, so the cost is independent of the size of the Dirichlet parameter space. Because $\tilde{p}$ depends on q and the lagged r_t is self-consistent only at convergence, F_T is monitored rather than assumed to descend. The count update is the plain conjugate one: exact stationarity of F_T in q_{θ_A} adds only a centered score that shifts no count mass (Remark 2). As $q_{\theta_A}, q_{\theta_B} \to \delta_{\theta^*}$ the weights vanish, $m \to p(\cdot \mid \theta^*, \theta_c)$, and (17)–(18) reduce to the fixed-parameter EFE kernels of [13, 14] with their enumerable θ read as θ_c; $K = 1$ removes the context (Section 5).

5 Experiments

We test whether the information-gain corrections of (16) produce exploration that a preference-driven planner cannot, on two grid POMDPs where knowing the model changes the route and exploration is survivable. Everything outside the objective is shared: the conjugate graph of Section 4 with $K = 1$, the exact filter and count update of Algorithm 1, a terminal preference on the goal, and identical horizons, sweeps and flat Dirichlet priors on the unknown block (details in Appendix C).[1]

Methods. **SP-EFE** is the full scheme with kernels (17)–(18). **SP-Pref-Plan** keeps only the planning correction $r_t(u_t \mid x_{t-1})$ (uniform channels, $\Delta_A = \Delta_B = 0$). **SP-Pref** drops the planning correction too. Both are blind to m_A in planning, since their observation-to-state message is $\sum_y m_A(y \mid x) = 1$, and reach an unknown transition only through posterior-mean *optimism*. Four ablations drop terms of (16) from SP-EFE: SP-EFE$^{-\mathrm{dyn}}$ (no dynamics channel), SP-EFE$^{-\Delta_B}$, and SP-EFE$^{-\Delta_A-\Delta_B}$ with and without the channel (ambiguity, no novelty). **Random** acts uniformly while learning; **EFE** enumerates the 4^H policies under the exact (7).

Learning the observation model (fog-walls). In the fog-walls maze (Figure 2a) the direct 4-step route crosses a drift cell and an ambiguous fog block, whereas the 8-step top row identifies every state; walls make mislocalization cost bumps rather than the episode. The top row pays off only for a planner that *knows* it is unambiguous: with A, B known, SP-EFE takes it on 30/30 seeds, whereas SP-Pref-Plan, whose route is A-independent, enters the fog and scores 18/30 (Appendix C.2). When learning A (Figure 2b–c, Table 1), SP-EFE reaches a goal rate of 1.00 from episode ≈30 on every seed, while SP-Pref-Plan plateaus at 0.81 and SP-Pref at 0.48 with 1.4–1.9× the matrix error: by construction, learning cannot change their route. The curve of SP-EFE is a novelty phase

[1] Code: https://github.com/skoghoern/sp-efe-parameter-learning

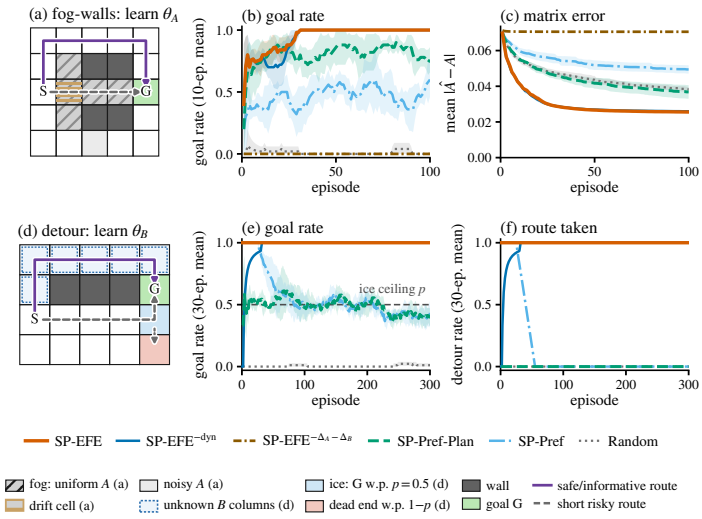


Fig. 2. Top (a–c): learning θ_A on fog-walls (B known, $H = 8$, 5 seeds); **bottom (d–f): learning θ_B** on the unmapped detour (identity A, $H = 7$, $T = 8$, 3 seeds). Curves are trailing means with a ± 1 std band over seeds; the glyph legend gives the cell and route vocabulary of (a) and (d).

followed by a switch: under the flat prior every cell is equally novel, so the agent enters the fog; once $\mathbb{H}[m_A(\cdot \mid x)]$ is $\log 5$ there and near 0 elsewhere, ambiguity reroutes the plan along a deterministic row and the decayed Δ_A no longer pulls it back.

Learning the transition model (unmapped detour). The unmapped detour (Figure 2d) caps a planner that never learns at $p = 0.5$, and optimism is priced out by design: a flat column spreads the successor over all $N = 20$ cells, which reaches the goal within the horizon with probability $0.265 < p$, so under its own objective SP-Pref-Plan *should* gamble. Table 1 reports three scenarios. With B known, every SP-EFE variant and SP-Pref-Plan take the detour on 300/300 episodes, so learning is *sufficient*; SP-Pref gambles even with a perfect model (Appendix C.2). With B frozen at the prior, SP-Pref-Plan sits at $146 \approx p$ and never opens a door, so learning is *necessary*. When learning B, SP-EFE takes the detour from episode 1 and never leaves it, whereas SP-Pref-Plan gambles on all 300 episodes and m_B ends equal to its prior. SP-Pref walks the detour only while the smeared doors still pay its action-entropy bonus, then gambles for the remaining 275 episodes: a better model removes the only term that steered it to the safe route. Enumerated EFE takes the detour on 100/100 episodes, so the exploration is a property of the objective, not of the message passing.

Ablation. The lower block of Table 1 removes terms of (16). *Novelty is the drive.* Without Δ_A, the channels-only planner keeps the ambiguity penalty but has nothing that rewards resolving it; from a flat prior every unvisited cell is

Table 1. Results and ablation. Fog-walls, learning θ_A (100 episodes, 5 seeds): goals, goals in the last 20, final matrix error. Unmapped detour, learning θ_B (300 episodes, 3 seeds): goals with B known, frozen at the prior, or learned, and learned-B goals via the detour. Lower block: ablations of (16). †Entropy-following: with frozen flat doors, the most uncertain column and the detour coincide.

	fog-walls, learn θ_A			unmapped detour, learn θ_B			
Planner	goals/100	last 20	err_A	known	frozen	learned	detour
SP-EFE	$\mathbf{95.8 \pm 0.7}$	**20.0**	**0.026**	**300**	300†	**300**	**300**
SP-Pref-Plan	81.4 ± 4.8	16.2	0.037	300	146	146 ± 7	0
SP-Pref	47.4 ± 3.1	9.6	0.049	146	300†	158 ± 8	25
Random	1.0 ± 0.9	0.2	0.038	1	1	1.3 ± 0.5	0
SP-EFE$^{-\mathrm{dyn}}$	94.4 ± 2.0	20.0	0.026	300	0	298	298
SP-EFE$^{-\Delta_B}$	94.4 ± 2.2	20.0	0.025	300	300†	300	300
SP-EFE$^{-\Delta_A-\Delta_B}$	0.0 ± 0.0	0.0	0.071	300	300†	300	300
SP-EFE$^{-\Delta_A-\Delta_B-\mathrm{dyn}}$	0.0 ± 0.0	0.0	0.071	300	146	146 ± 7	0

maximally ambiguous ($\log 25$ nats per step against a preference of 4), and the agent bumps into the wall beside the start on all 500 episodes. Without Δ_B, the learned-B score is carried by the dynamics channel, which the frozen column exposes as entropy-following. *The dynamics channel is dispensable.* Removing it changes nothing on fog-walls and costs two episodes on the detour; at its fixed point the channel cancels m_B on its support, so a smeared learned m_B loses the exploitation signal.

6 Conclusion

This paper extended entropy-corrected EFE-based planning [14] from inference over a finite set of known models to conjugate *learning* of the observation and transition matrices, as the minimization of a single variational objective. Frozen per-factor copies keep every factor belief conjugate and preserve the parameter–datum joint that mean-field destroys, so the parameter side collapses into predictive likelihoods while the information gains survive as kernel weights. On two grid POMDPs those weights are the drive: the scheme learns θ_A or θ_B and settles on the safe route where SP-MP with the same filter and count update plateaus or gambles.

Limitations and future work. The empirical advantage has conditions, all visible in the designs of Section 5: exploration must be survivable, the unknown region small enough for novelty to satiate, and model knowledge must change the route, since greedy optimism already wins on a unique-path task with a hard deadline. The objective is non-convex and the kernels use lagged factors, so damped descent need not be monotone and F_T measures deviations without certifying convergence. Future work includes descent-guided damping, score-corrected learning, and sophisticated planning beyond the frozen parameter priors.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Attias, H.: Planning by probabilistic inference. In: International Workshop on Artificial Intelligence and Statistics (AISTATS). pp. 9–16. PMLR (2003)

[2] Champion, T., Da Costa, L., Bowman, H., Grześ, M.: Branching time active inference: The theory and its generality. Neural Networks **151**, 295–316 (2022). https://doi.org/10.1016/j.neunet.2022.03.036

[3] Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., Friston, K.: Active inference on discrete state-spaces: A synthesis. Journal of Mathematical Psychology **99**, 102447 (Dec 2020). https://doi.org/10.1016/j.jmp.2020.102447, http://dx.doi.org/10.1016/j.jmp.2020.102447

[4] Friston, K., Da Costa, L., Hafner, D., Hesp, C., Parr, T.: Sophisticated inference. Neural Computation **33**(3), 713–763 (2021). https://doi.org/10.1162/neco_a_01351

[5] Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., Pezzulo, G.: Active inference and learning. Neuroscience & Biobehavioral Reviews **68**, 862–879 (2016). https://doi.org/10.1016/j.neubiorev.2016.06.022

[6] Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., Pezzulo, G.: Active inference and epistemic value. Cognitive Neuroscience **6**(4), 187–214 (2015). https://doi.org/10.1080/17588928.2015.1020053

[7] Kappen, H.J., Gómez, V., Opper, M.: Optimal control as a graphical model inference problem. Machine Learning **87**(2), 159–182 (2012). https://doi.org/10.1007/s10994-012-5278-7

[8] van de Laar, T., Koudahl, M., de Vries, B.: Realizing synthetic active inference agents, part ii: Variational message updates. Neural Computation **37**(1), 38–75 (2024)

[9] Levine, S.: Reinforcement learning and control as probabilistic inference: Tutorial and review. arXiv preprint arXiv:1805.00909 (2018)

[10] Lázaro-Gredilla, M., Ku, L.Y., Murphy, K.P., George, D.: What type of inference is planning? In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024) (2024)

[11] Nuijten, W.W.L., van de Laar, T., de Vries, B.: Expected free energy-based planning as variational inference. Transactions on Machine Learning Research (2026), https://openreview.net/pdf?id=Kzm8I1oS1s

[12] Nuijten, W.W.L., Lukashchuk, M.: Active inference is a subtype of variational inference. In: EurIPS Workshop on Epistemic Intelligence in Machine Learning (nov 2025)

[13] Nuijten, W.W.L., Lukashchuk, M., van de Laar, T., de Vries, B.: A message passing realization of expected free energy minimization. In: International Workshop on Active Inference (IWAI) (oct 2025)

[14] Nuijten, W.W.L., Lukashchuk, M., van de Laar, T., de Vries, B.: What type of inference is active inference? In: Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI) (2026)

[15] Parr, T., Friston, K.J.: Generalised free energy and active inference. Biological Cybernetics **113**(5), 495–513 (2019). https://doi.org/10.1007/s00422-019-00805-w

[16] Paul, A., Sajid, N., Da Costa, L., Razi, A.: On efficient computation in active inference. Expert Systems with Applications **253**, 124315 (2024). https://doi.org/10.1016/j.eswa.2024.124315

[17] Toussaint, M.: Robot trajectory optimization using approximate inference. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML). pp. 1049–1056 (2009). https://doi.org/10.1145/1553374.1553508

[18] Şenöz, I., van de Laar, T., Bagaev, D., de Vries, B.: Variational message passing and local constraint manipulation in factor graphs. Entropy **23**(7), 807 (jun 2021). https://doi.org/10.3390/e23070807, https://doi.org/10.3390/e23070807

A The Objective: Proofs and Constructions

Notation: $\theta = (\theta_A, \theta_B, \theta_c)$, $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}, \theta)$, $\mathbb{E}_q = \mathbb{E}_{q(\boldsymbol{z})}$; $h[q(\cdot \mid x)]$ is the entropy of a conditional as a function of x and $\mathbb{H}[q(\cdot \mid x)] = \mathbb{E}_{q(x)} h[q(\cdot \mid x)]$ the conditional-entropy scalar; $\mathrm{Dir}(\alpha)$ with $\alpha_0 = \sum_k \alpha_k$ has density $\prod_k \theta_k^{\alpha_k - 1}/\mathrm{B}(\alpha)$. Time indices are suppressed where no confusion arises.

A.1 Epistemic Priors Recover the EFE

Lemma 1 (Prior contributions). *For the priors* (8)–(11) *and any joint* q *possessing the conditionals they refer to,*

$$-\mathbb{E}_q[\log \tilde{p}(u_t)] = -\mathbb{H}[q(x_t \mid x_{t-1}, u_t)], \tag{19}$$

$$-\mathbb{E}_q[\log \tilde{p}(x_t)] = +\mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)], \tag{20}$$

$$-\mathbb{E}_q[\log \tilde{p}(y_t, x_t)] = -I(\theta_A, \theta_c; y_t \mid x_t) = -\mathcal{I}_{c,t} - \mathcal{I}_{A,t}, \tag{21}$$

$$-\mathbb{E}_q[\log \tilde{p}(x_t, x_{t-1}, u_t, \theta_c)] = -I(\theta_B; x_t, u_t \mid x_{t-1}, \theta_c) = -\mathcal{I}_{B,t}, \tag{22}$$

up to the normalizing constants of the priors.

Proof. Each prior is the exponential of a functional of the variables it is indexed by; taking $-\log$ and the outer expectation over those variables gives the stated quantity by definition. For (19), $\mathbb{E}_{q(u_t)}\mathbb{E}_{q(x_t, x_{t-1} \mid u_t)}[\log q(x_t \mid x_{t-1}, u_t)] = -\mathbb{H}[q(x_t \mid x_{t-1}, u_t)]$; for (20), $\mathbb{E}_{q(x_t)}\mathbb{E}_{q(\theta_A, \theta_c \mid x_t)} h[q(y_t \mid x_t, \theta_A, \theta_c)] = \mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)]$; for (21), $\mathbb{E}_{q(y_t, x_t)} D_{\mathrm{KL}}[q(\theta_A, \theta_c \mid y_t, x_t) \,\|\, q(\theta_A, \theta_c \mid x_t)]$ is the conditional mutual information $I(\theta_A, \theta_c; y_t \mid x_t)$ written as an expected divergence between posterior and cavity, and the chain rule $I(\theta_A, \theta_c; y_t \mid x_t) = I(\theta_c; y_t \mid x_t) + I(\theta_A; y_t \mid x_t, \theta_c)$ gives the split into salience and novelty of Section 4.1; likewise (22) with (x_t, u_t) as the datum and (x_{t-1}, θ_c) as the conditioning variables. □

Proof (of Corollary 1). Split $\log \tilde{p}$ into the unbiased model, the goals and the epistemic priors,

$$\log \tilde{p} = \log p + \sum_{t=1}^{T} \log \hat{p}_x(x_t)\, \hat{p}_y(y_t) + \sum_{t=1}^{T} \log \big[\tilde{p}(u_t)\, \tilde{p}(x_t)\, \tilde{p}(y_t, x_t)\, \tilde{p}(x_t, x_{t-1}, u_t, \theta_c)\big]. \tag{23}$$

By Lemma 1,

$$\begin{aligned} F_{\tilde{p}}[q] = \underbrace{\mathbb{E}_q\Big[\log \frac{q}{p}\Big]}_{B[q]} + \sum_{t=1}^{T} \Big(&- \mathbb{E}_q\big[\log \hat{p}_x(x_t) \hat{p}_y(y_t)\big] - \mathbb{H}[q(x_t \mid x_{t-1}, u_t)] \\ &+ \mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)] - \mathcal{I}_{c,t} - \mathcal{I}_{A,t} - \mathcal{I}_{B,t}\Big). \end{aligned} \tag{24}$$

Expanding (7) and taking $\mathbb{E}_{q(u_t)}$,

$$\begin{aligned} \mathbb{E}_{q(u_t)}[G(u_t)] = &- \mathbb{E}_q\big[\log \hat{p}_x(x_t) \hat{p}_y(y_t)\big] - \mathbb{H}[q(x_t \mid x_{t-1}, u_t)] \\ &+ \mathbb{E}_q[-\log p(y_t \mid x_t, \theta_A, \theta_c)] - \mathcal{I}_{c,t} - \mathcal{I}_{A,t} - \mathcal{I}_{B,t}, \end{aligned} \tag{25}$$

which coincides with the t-th summand above whenever $q(y_t \mid x_t, \theta_A, \theta_c) = p(y_t \mid x_t, \theta_A, \theta_c)$. For the factor beliefs of Section 4.3, $q_{\mathrm{obs},t}(y_t \mid x_t, \theta_A, \theta_c) \propto p(y_t \mid x_t, \theta_A, \theta_c)\, \hat{p}_y(y_t)$, so the identity is exact for uniform $\hat{p}_y$ (the setting of Section 5); in general the ambiguity term of G is to be read as $\mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)]$, the entropy of the agent's own predictive observation channel, as in [14]. Summing over t gives the Corollary. □

A.2 Entropy Corrections

Adding the planning correction $+\mathbb{H}[q(u_t \mid x_{t-1})]$ [10] to the sum of (19)–(22) gives the raw per-step correction

$$\Delta F_t = \mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)] - \mathcal{I}_{c,t} - \mathcal{I}_{A,t} - \mathbb{H}[q(x_t \mid x_{t-1}, u_t)] - \mathcal{I}_{B,t} + \mathbb{H}[q(u_t \mid x_{t-1})]. \tag{26}$$

The identity $\mathbb{H}[q(y_t \mid x_t, \theta_A, \theta_c)] = \mathbb{H}[q(y_t \mid x_t)] - I(\theta_A, \theta_c; y_t \mid x_t) = \mathbb{H}[q(y_t \mid x_t)] - \mathcal{I}_{c,t} - \mathcal{I}_{A,t}$ turns (26) into (12): the ambiguity prior contributes one unit of each observation-side gain through its θ-coupling, the novelty prior a second. Under any family with $q(y_t, x_t, \theta_c, \theta_A) = q(y_t, x_t, \theta_c)\, q(\theta_A)$ the Dirichlet gain $\mathcal{I}_{A,t}$ vanishes identically and prior (10) retains only its salience part; the same holds for $\mathcal{I}_{B,t}$ and (11).

A.3 The Conjugate Factor Belief

Lemma 2 (Entrywise conjugacy). *For one column $\alpha = \alpha_A(\cdot \mid x, \theta_c)$ and any entry y, $\theta_y \operatorname{Dir}(\theta \mid \alpha) = \dfrac{\alpha_y}{\alpha_0} \operatorname{Dir}(\theta \mid \alpha + e_y)$.*

Proof. $\theta_y \prod_k \theta_k^{\alpha_k - 1} / \mathrm{B}(\alpha) = \prod_k \theta_k^{\alpha_k + [k=y] - 1} / \mathrm{B}(\alpha) = \dfrac{\mathrm{B}(\alpha + e_y)}{\mathrm{B}(\alpha)} \operatorname{Dir}(\theta \mid \alpha + e_y)$ and $\dfrac{\mathrm{B}(\alpha + e_y)}{\mathrm{B}(\alpha)} = \dfrac{\Gamma(\alpha_y + 1)\, \Gamma(\alpha_0)}{\Gamma(\alpha_y)\, \Gamma(\alpha_0 + 1)} = \dfrac{\alpha_y}{\alpha_0}$. Other columns are untouched. □

Applied, for each value of θ_c, to the factor belief of Section 4.3,

$$q_{\mathrm{obs},t}(y_t, x_t, \theta_c, \theta_A^{(t)}) \propto p(y_t \mid x_t, \theta_A^{(t)}, \theta_c)\, \mu_{y_t}(y_t)\, \mu_{x_t}(x_t)\, \mu_{\theta_c}(\theta_c) \operatorname{Dir}(\theta_A^{(t)} \mid \alpha_A),$$

this gives (13). The identity needs two things: the message into the copy must be free of (y, x, θ_c)—true because the copy has degree two and its other neighbor is the prior node—and the kernel must be a monomial in θ_A for each value of θ_c. A θ_A-conditioned channel inside the kernel, such as $r(y \mid x, \theta_A, \theta_c)^2$ with the shared parameter of [14] read as θ_A, has a normalizer that is not polynomial in θ_A and would destroy the identity; this is why the θ_A-epistemic content enters the kernel as a scalar weight per discrete entry, while the finite θ_c retains the channel form (Proposition 2).

Lemma 3 (Elimination of the θ_A-conditional). *Let the observation region carry the belief $q_{yx,t}(y, x, \theta_c)\, \nu_t(\theta_A \mid y, x, \theta_c)$ for an arbitrary normalized ν_t, and*

let the copy's prior region carry the induced marginal $\sum_{y,x,\theta_c} q_{yx,t}\nu_t$. *Their joint contribution to the Bethe free energy of the frozen graph is*

$$-\mathbb{H}[q_{yx,t}]-\mathbb{E}_{q_{yx,t}}[\log m_A]+\sum_{y,x,\theta_c} q_{yx,t}(y,x,\theta_c)\, D_{\mathrm{KL}}\big[\nu_t(\cdot \mid y,x,\theta_c)\,\big\|\,\mathrm{Dir}(\alpha_A+e_{yx\theta_c})\big], \tag{27}$$

minimized uniquely at $\nu_t = \mathrm{Dir}(\alpha_A + e_{yx\theta_c}) = q_{\mathrm{obs},t}(\theta_A \mid x,y,\theta_c)$.

Proof. The copy has degree two, so its region term $D_{\mathrm{KL}}[q(\theta_A^{(t)}) \,\|\, \mathrm{Dir}(\alpha_A)]$ and its counting term $+\mathbb{H}[q(\theta_A^{(t)})]$ combine to $-\mathbb{E}_{q(\theta_A^{(t)})}[\log \mathrm{Dir}(\theta_A \mid \alpha_A)]$. Adding the observation region $D_{\mathrm{KL}}[q_{yx,t}\nu_t \,\|\, p(y \mid x,\theta_A,\theta_c)]$ and using $q(\theta_A^{(t)}) = \sum_{y,x,\theta_c} q_{yx,t}\nu_t$,

$$-\mathbb{H}[q_{yx,t}] + \sum_{y,x,\theta_c} q_{yx,t}\, \mathbb{E}_{\nu_t}\Big[\log \frac{\nu_t(\theta_A \mid y,x,\theta_c)}{\mathrm{Dir}(\theta_A \mid \alpha_A)\, p(y \mid x,\theta_A,\theta_c)}\Big],$$

and Lemma 2 rewrites the denominator as $m_A(y \mid x,\theta_c)\,\mathrm{Dir}(\theta_A \mid \alpha_A+e_{yx\theta_c})$. □

A coordinate whose optimum is a closed-form function of the others is substituted out; it is also not storable, θ_A being continuous. Two consequences. First, the parameter energy and entropy of the region collapse jointly into the *predictive* energy $-\log m_A$; nothing θ_A-dependent survives in the base free energy (used in Appendix A.5). Second, the corrections (12) are *evaluated* at $\nu_t = \mathrm{Dir}(\alpha_A+e_{yx\theta_c})$ and not re-minimized over ν_t: read as a functional of a free ν_t, $-2\mathcal{I}_{A,t}$ would reward arbitrary coupling up to its ceiling $\mathbb{H}_q[y_t \mid x_t,\theta_c]$. The dynamics side is identical with columns indexed by (x_{t-1},u_t,θ_c). The finite θ_c is treated differently: a free θ_c-axis on the discrete beliefs costs a factor K and is kept, so its conditionals are exact and no projection is needed.

The step-wise copies are what make this consistent. A single shared θ_A would impose the calibration constraint $\sum_{y,x,\theta_c} q_{\mathrm{obs},t} = q(\theta_A)$ for every t, but $\sum_{y,x,\theta_c} q_{\mathrm{obs},t} = \mathrm{Dir}(\alpha_A)\sum_{y,x,\theta_c} q_{yx,t}(y,x,\theta_c)\,\theta_A(y \mid x,\theta_c)/m_A(y \mid x,\theta_c)$ equals $\mathrm{Dir}(\alpha_A)$ only when $q_{yx,t}(y \mid x,\theta_c) = m_A(y \mid x,\theta_c)$, i.e. when planning leaves the block conditionals at their predictive values; enforcing it would write $\sim T$ imagined counts into α_A per sweep and eliminate novelty inside the planning loop. With step-wise copies the constraint holds trivially and the approximation moves out of the constraint set into the planning-time model.

A.4 The Cavity and Its Projection

Fix a column (x,θ_c), write $r_y := q_{yx,t}(y \mid x,\theta_c)$, $p_y := \mathrm{Dir}(\alpha + e_y)$, and let $\bar{q} := \sum_y r_y\, p_y$ be the exact cavity (14). The projection $\tilde{q} := \mathrm{Dir}(\alpha + r)$ has the following properties.

(i) *Mean matching.* $\mathbb{E}_{\bar{q}}[\theta_k] = \sum_y r_y \frac{\alpha_k+[k=y]}{\alpha_0+1} = \frac{\alpha_k+r_k}{\alpha_0+1} = \mathbb{E}_{\tilde{q}}[\theta_k]$.

(ii) *Geometric mixture.* $\prod_y p_y^{r_y} \propto \prod_k \theta_k^{\sum_y r_y(\alpha_k+[k=y]-1)} = \prod_k \theta_k^{\alpha_k+r_k-1} \propto \tilde{q}$: the projection averages the branches' natural parameters. Equivalently $\tilde{q} = \arg\min_{\tilde{q}'} \sum_y r_y D_{\mathrm{KL}}[\tilde{q}' \,\|\, p_y]$, whereas $\bar{q} = \arg\min_{\tilde{q}'} \sum_y r_y D_{\mathrm{KL}}[p_y \,\|\, \tilde{q}']$.

(iii) *Structured VMP.* $\tilde{q}$ is what variational message passing returns on the θ-edge under $q(y, \theta \mid x) = q(y \mid x)\, q(\theta \mid x)$: prior times $\exp \mathbb{E}_{q(y|x)}[\log \theta_y] = \prod_y \theta_y^{r_y}$ [18]; the sum-product message $\sum_y r_y \theta_y$ gives the arithmetic mixture $\bar{q}$ instead.

(iv) *Not the M-projection.* $\arg\min_{\tilde{q}' \in \mathrm{Dir}} D_{\mathrm{KL}}[\bar{q} \,\|\, \tilde{q}']$ would match $\mathbb{E}[\log \theta_k]$. The mixture carries between-branch variance that a single Dirichlet drops, $\mathbb{E}_{\bar{q}}[\theta_k^2] - \mathbb{E}_{\tilde{q}}[\theta_k^2] = \frac{r_k(1-r_k)}{(\alpha_0+1)(\alpha_0+2)}$, so $\tilde{q}$ is slightly overconfident and $\Delta_{A,t}$ slightly overstated, with relative error $O(1/\alpha_0)$ that vanishes as counts grow.

(v) *Exactness and closure.* $\tilde{q} = \bar{q}$ whenever r is one-hot; $\tilde{q}$ stays Dirichlet under linear damping of r; and $D_{\mathrm{KL}}[\mathrm{Dir} \,\|\, \mathrm{Dir}]$ is closed-form in digamma and log-gamma functions.

The projection is used for the weight arrays only; the discrete messages close through m_A, m_B without it. Its one side effect on the stationarity conditions is quantified in Lemma 5.

A.5 Proof of Proposition 1

Degrees and Bethe free energy. On the frozen graph, y_t, u_t, x_0 and every copy $\theta_A^{(t)}, \theta_B^{(t)}$ have one non-singleton neighbor besides their prior or goal node; x_t for $1 \le t < T$ has three $(f_{\mathrm{obs},t}, f_{\mathrm{dyn},t}, f_{\mathrm{dyn},t+1})$, x_T two, and θ_c all $2T$. For a variable with k non-singleton neighbors and singleton factor f_i, the singleton region and the counting term combine to $(k-1)\,\mathbb{H}[q_i] - \mathbb{E}_{q_i}[\log f_i]$. Hence

$$
\begin{aligned}
F_{\hat{p}}[q] = \sum_{t=1}^{T} \Big(& D_{\mathrm{KL}}\big[q_{\mathrm{obs},t} \,\|\, p(y_t \mid x_t, \theta_A^{(t)}, \theta_c)\big] - \mathbb{E}_{q(\theta_A^{(t)})}\big[\log \mathrm{Dir}(\theta_A^{(t)} \mid \alpha_A)\big] \\
& + D_{\mathrm{KL}}\big[q_{\mathrm{dyn},t} \,\|\, p(x_t \mid x_{t-1}, u_t, \theta_B^{(t)}, \theta_c)\big] - \mathbb{E}_{q(\theta_B^{(t)})}\big[\log \mathrm{Dir}(\theta_B^{(t)} \mid \alpha_B)\big] \\
& - \mathbb{E}_q\big[\log p(u_t)\, \hat{p}_y(y_t)\, \hat{p}_x(x_t)\big] \Big) \\
& + 2 \sum_{t=1}^{T-1} \mathbb{H}[q(x_t)] + \mathbb{H}[q(x_T)] + (2T-1)\,\mathbb{H}[q(\theta_c)] - \mathbb{E}_q[\log p(x_0)\, p(\theta_c)],
\end{aligned}
\tag{28}
$$

and the same degree count gives the exponents of the Bethe joint of Section 4.4. Under the planning convention that $p(\theta_c)$ is the current context belief, the θ_c terms are constants during planning.

(i) Base. Substitute (13), i.e. $\nu_t = \mathrm{Dir}(\alpha_A + e_{yx\theta_c})$, into Lemma 3: the divergence term vanishes and the first two terms of (28) equal $-\mathbb{H}[q_{yx,t}] - \mathbb{E}_{q_{yx,t}}[\log m_A]$;

likewise on the dynamics side with m_B. Hence

$$F_{\hat{p}}[q] = \sum_{t=1}^{T} \Big(-\mathbb{H}[q_{yx,t}] - \mathbb{E}_{q_{yx,t}}[\log m_A] - \mathbb{H}[q_{\mathrm{trip},t}] \\ - \mathbb{E}_{q_{\mathrm{trip},t}}[\log m_B] - \mathbb{E}_q\big[\log p(u_t)\,\hat{p}_y(y_t)\,\hat{p}_x(x_t)\big]\Big) - \mathbb{E}_q[\log p(x_0)\,p(\theta_c)] \\ + 2\textstyle\sum_{t=1}^{T-1}\mathbb{H}[q(x_t)] + \mathbb{H}[q(x_T)] + (2T-1)\,\mathbb{H}[q(\theta_c)]. \tag{29}$$

Per entry the cancellation reads

$$-h\big[\mathrm{Dir}(\alpha_A + e_{yx\theta_c})\big] - \mathbb{E}_{\mathrm{Dir}(\alpha_A+e_{yx\theta_c})}\big[\log p(y \mid x, \theta_A, \theta_c) + \log \mathrm{Dir}(\theta_A \mid \alpha_A)\big] \\ = -\log m_A(y \mid x, \theta_c), \tag{30}$$

which is Bayes' rule $\log \mathrm{Dir}(\alpha_A + e_{yx\theta_c}) = \log \mathrm{Dir}(\alpha_A) + \log p - \log m_A$ integrated against itself. Because the copy's prior node carries the current belief $\mathrm{Dir}(\alpha_A)$, there is no prior-mismatch residual $\mathbb{E}[\log q(\theta_A)/p(\theta_A)]$ to dispose of.

(ii) Corrections. Each copy lies in exactly one region of the Bethe joint, so the marginal of q over $(y_t, x_t, \theta_c, \theta_A^{(t)})$ is $q_{\mathrm{obs},t}$ and over $(x_t, x_{t-1}, u_t, \theta_c, \theta_B^{(t)})$ is $q_{\mathrm{dyn},t}$; on the chain the Bethe joint is a proper distribution with these marginals. Its θ_A, θ_B-free conditionals are the block conditionals, $q(y_t \mid x_t) = q_{yx,t}(y_t \mid x_t)$, $q(\theta_c \mid x_t, y_t) = q_{yx,t}(\theta_c \mid x_t, y_t)$, $q(x_t \mid x_{t-1}, u_t) = q_{\mathrm{trip},t}(x_t \mid x_{t-1}, u_t)$, $q(u_t \mid x_{t-1}) = q_{\mathrm{trip},t}(u_t \mid x_{t-1})$; its θ_A-conditionals are $q_{\mathrm{obs},t}(\theta_A \mid x_t, y_t, \theta_c) = \mathrm{Dir}(\alpha_A + e_{y_t x_t \theta_c})$ and the cavity (14). Therefore $I_q(\theta_c; y_t \mid x_t) = I_{q_{yx,t}}(y_t; \theta_c \mid x_t)$, $I_q(\theta_A^{(t)}; y_t \mid x_t, \theta_c) = \mathbb{E}_{q_{yx,t}} D_{\mathrm{KL}}[q_{\mathrm{obs},t}(\theta_A \mid x_t, y_t, \theta_c) \,\|\, q_{\mathrm{obs},t}(\theta_A \mid x_t, \theta_c)] = \mathbb{E}_{q_{yx,t}}[\Delta_{A,t}]$, likewise for θ_B, and (12) holds with every term read off $q_{yx,t}$ and $q_{\mathrm{trip},t}$. Nothing is double-counted with (i): the information-gain mass in the θ-coupled energy is cancelled by (30) and never reaches ΔF_t.

(iii) Assembly. Adding (i) and (ii) gives (16), a functional of the discrete coordinates of Section 4.3 and the fixed counts α_A, α_B only. □

Remark 1 (Limits). At $q(\theta_A), q(\theta_B) = \delta_{\theta^*}$ every Dirichlet posterior and cavity is δ_{θ^*}, $\Delta_A = \Delta_B = 0$, $m \to p(\cdot \mid \theta^*, \theta_c)$, and (16) is the fixed-parameter EFE-planning objective of [14] with their θ read as θ_c. Deleting the correction line leaves belief propagation on the posterior predictive likelihoods. Both limits are approached continuously in the concentrations.

B Message-Passing Derivation

We derive the stationarity conditions of the planning objective (16) and solve them as message passing, following Appendix D of [14] with two structural differences: the Dirichlet copies are integrated out in closed form, so there are no θ_A, θ_B-messages, and the channels are conditionals of q, so there are no channel variables or constraints. The context θ_c is a discrete variable of the graph and is treated exactly as the shared θ of [14].

B.1 Lagrangian

At $T{=}1$ the coordinates are $q_{yx}(y_1, x_1, \theta_c)$, $q_{\mathrm{trip}}(x_1, x_0, u_1, \theta_c)$ and the singletons $q_{x_0}, q_{x_1}, q_{y_1}, q_{u_1}, q_{\theta_c}$; the parameter beliefs $\mathrm{Dir}(\alpha_A), \mathrm{Dir}(\alpha_B)$ are fixed. With multipliers for normalization of the two blocks and for the seven factor-to-singleton consistency constraints,

$$\begin{aligned}\mathcal{L}[q,\lambda] = F_1[q] &+ \lambda_{yx}\Big(\textstyle\sum q_{yx} - 1\Big) + \lambda_{\mathrm{trip}}\Big(\textstyle\sum q_{\mathrm{trip}} - 1\Big) \\ &+ \textstyle\sum_{y_1} \lambda_{y_1}\Big(\sum_{x_1,\theta_c} q_{yx} - q_{y_1}\Big) + \sum_{x_1} \lambda^{\mathrm{obs}}_{x_1}\Big(\sum_{y_1,\theta_c} q_{yx} - q_{x_1}\Big) \\ &+ \textstyle\sum_{\theta_c} \lambda^{\mathrm{obs}}_{\theta_c}\Big(\sum_{y_1,x_1} q_{yx} - q_{\theta_c}\Big) + \sum_{x_1} \lambda^{\mathrm{dyn}}_{x_1}\Big(\sum_{x_0,u_1,\theta_c} q_{\mathrm{trip}} - q_{x_1}\Big) \\ &+ \textstyle\sum_{x_0} \lambda_{x_0}\Big(\sum_{x_1,u_1,\theta_c} q_{\mathrm{trip}} - q_{x_0}\Big) + \sum_{u_1} \lambda_{u_1}\Big(\sum_{x_1,x_0,\theta_c} q_{\mathrm{trip}} - q_{u_1}\Big) \\ &+ \textstyle\sum_{\theta_c} \lambda^{\mathrm{dyn}}_{\theta_c}\Big(\sum_{x_1,x_0,u_1} q_{\mathrm{trip}} - q_{\theta_c}\Big),\end{aligned} \tag{31}$$

where each consistency multiplier is a function of its variable, $\lambda_{y_1} = \lambda_{y_1}(y_1)$ etc. Relative to [14] (their (D.3)–(D.7)), the four channel multipliers are absent and the θ-consistency multipliers refer to θ_c only. The corrections in F_1 are functionals of the coordinates through (13)–(14), cavities included.

B.2 Stationarity Conditions

Two lemmas make the variation of the corrections local.

Lemma 4 (Score lemma). *For a normalized joint $B[q](a,b)$ depending on a coordinate q, and any of its conditionals,*

$$\frac{\delta}{\delta q}\,\mathbb{H}\big[B(a \mid b)\big] = -\int \frac{\delta B(a,b)}{\delta q}\,\log B(a \mid b)\,da\,db, \tag{32}$$

i.e. entropies of conditionals—and hence conditional mutual informations—differentiate as if the logarithm were frozen.

Proof. $\mathbb{H}[B(a \mid b)] = -\int B(a,b)\log\frac{B(a,b)}{B(b)}$. The product-rule remainder is $-\int B(a,b)\big[\frac{\delta B(a,b)/\delta q}{B(a,b)} - \frac{\delta B(b)/\delta q}{B(b)}\big] = -\int \frac{\delta B(a,b)}{\delta q} + \int \frac{\delta B(b)}{\delta q} = 0$, since $\int_a B(a,b) = B(b)$ for every q. □

Lemma 5 (Cavity back-reaction vanishes). *With $\beta(\theta \mid x, y, \theta_c) := q_{\mathrm{obs},t}(\theta_A \mid x, y, \theta_c)$ independent of $q_{yx,t}$ and the exact cavity $\bar\beta(\theta \mid x, \theta_c) = \sum_y q_{yx,t}(y \mid x, \theta_c)\,\beta(\theta \mid x, y, \theta_c)$,*

$$\frac{\delta\,\mathcal{I}_{A,t}}{\delta q_{yx,t}(y,x,\theta_c)} = \Delta_{A,t}(x,y,\theta_c) = D_{\mathrm{KL}}\big[\beta(\cdot \mid x,y,\theta_c)\,\big\|\,\bar\beta(\cdot \mid x,\theta_c)\big], \tag{33}$$

with no contribution from the dependence of $\bar\beta$ on $q_{yx,t}$. Likewise $\delta\mathcal{I}_{B,t}/\delta q_{\mathrm{trip},t} = \Delta_{B,t}$.

Proof. Write

$$\mathcal{I}_{A,t} = \sum_{y,x,\theta_c} q_{yx,t}(y,x,\theta_c) \int \beta(\theta \mid x,y,\theta_c) \log \beta(\theta \mid x,y,\theta_c)\, d\theta - \sum_{y,x,\theta_c} q_{yx,t}(y,x,\theta_c) \int \beta(\theta \mid x,y,\theta_c) \log \bar{\beta}(\theta \mid x,\theta_c)\, d\theta.$$

The explicit derivative is $\Delta_{A,t}(x,y,\theta_c)$. Since $\sum_{y'} q_{yx,t}(y',x,\theta_c)\, \beta(\theta \mid x,y',\theta_c)$ equals $q(x,\theta_c)\, \bar{\beta}(\theta \mid x,\theta_c)$, the implicit derivative through $\bar{\beta}(\cdot \mid x,\theta_c)$ is

$$-q(x,\theta_c) \int \bar{\beta}(\theta \mid x,\theta_c) \frac{\delta \log \bar{\beta}(\theta \mid x,\theta_c)}{\delta q_{yx,t}(y,x,\theta_c)}\, d\theta = -q(x,\theta_c) \frac{\delta}{\delta q_{yx,t}(y,x,\theta_c)} \int \bar{\beta}(\theta \mid x,\theta_c)\, d\theta = 0,$$

because $\bar{\beta}$ is normalized for every q. □

Under the cavity projection of Section 4.3 the weight in the integral is the exact mixture while the logarithm is that of $\mathrm{Dir}(\alpha + \bar{n}_{x\theta_c})$, whose $\bar{n}$-derivative is $\log \theta_k - \psi(\alpha_k + \bar{n}_k) + \psi(\alpha_0 + 1)$; the residual is therefore proportional to $\mathbb{E}_{\bar{\beta}}[\log \theta_k] - \mathbb{E}_{\mathrm{Dir}(\alpha+\bar{n})}[\log \theta_k] = (1-r_k)\psi(\alpha_k) + r_k \psi(\alpha_k+1) - \psi(\alpha_k+r_k) = O(\alpha_k^{-2})$ by concavity of ψ, and is dropped. Only explicit appearances of q survive, which is why the updates stay local and the mixture structure of the cavity never enters the messages.

Proposition 3 (Observation block). *Varying* (31) *in* $q_{yx}(y_1,x_1,\theta_c)$*: the base* (29) *gives* $\log q_{yx} + 1 - \log m_A$*;* $+\mathbb{H}_{q_{yx}}[y_1 \mid x_1]$ *gives* $-\log q(y_1 \mid x_1)$ *and* $-2\mathcal{I}_c$ *gives* $-2\log[q(\theta_c \mid x_1,y_1)/q(\theta_c \mid x_1)]$ *(Lemma 4);* $-2\mathcal{I}_A$ *gives* $-2\Delta_A(x_1,y_1,\theta_c)$ *(Lemma 5); the multipliers give* $\lambda_{yx} + \lambda_{y_1} + \lambda_{x_1}^{\mathrm{obs}} + \lambda_{\theta_c}^{\mathrm{obs}}$*. Using* $q(\theta_c \mid x,y)/q(\theta_c \mid x) = q(y \mid x,\theta_c)/q(y \mid x)$*, at stationarity*

$$q^*_{yx}(y_1,x_1,\theta_c) \propto \underbrace{m_A(y_1 \mid x_1,\theta_c)}_{predictive} \underbrace{\frac{q^*(y_1 \mid x_1,\theta_c)^2}{q^*(y_1 \mid x_1)}}_{salience} \underbrace{e^{2\,\Delta_A(x_1,y_1,\theta_c)}}_{novelty \geq 1} e^{-\lambda^{\mathrm{obs}}}, \tag{34}$$

with $\lambda^{\mathrm{obs}} := \lambda_{y_1} + \lambda_{x_1}^{\mathrm{obs}} + \lambda_{\theta_c}^{\mathrm{obs}}$.

Proposition 4 (Dynamics block). *Varying in* $q_{\mathrm{trip}}(x_1,x_0,u_1,\theta_c)$*: the base gives* $\log q_{\mathrm{trip}} + 1 - \log m_B$*;* $-\mathbb{H}_q[x_1 \mid x_0,u_1]$ *gives* $+\log q(x_1 \mid x_0,u_1)$*;* $+\mathbb{H}_q[u_1 \mid x_0]$ *gives* $-\log q(u_1 \mid x_0)$*;* $-\mathcal{I}_B$ *gives* $-\Delta_B$*. At stationarity,*

$$q^*_{\mathrm{trip}}(x_1,x_0,u_1,\theta_c) \propto m_B(x_1 \mid x_0,u_1,\theta_c) \frac{q^*(u_1 \mid x_0)}{q^*(x_1 \mid x_0,u_1)} e^{\Delta_B(x_1,x_0,u_1,\theta_c)} e^{-\lambda^{\mathrm{dyn}}}, \tag{35}$$

with $\lambda^{\mathrm{dyn}} := \lambda_{x_1}^{\mathrm{dyn}} + \lambda_{x_0} + \lambda_{u_1} + \lambda_{\theta_c}^{\mathrm{dyn}}$.

Propositions 3–4 are Proposition 2: the kernels (13a)–(13b) of [14] with the point likelihood replaced by the posterior predictive likelihood and the

information-gain weights appended. Their observation kernel carries the ratio $q(y \mid x,\theta)^2/q(y \mid x)$ for the full θ; with $\theta = (\theta_A, \theta_c)$ it factorizes as

$$\begin{aligned} \frac{q(y \mid x,\theta_A,\theta_c)^2}{q(y \mid x)} &= \frac{q(y \mid x,\theta_c)^2}{q(y \mid x)} \cdot \Big[\frac{q(y \mid x,\theta_A,\theta_c)}{q(y \mid x,\theta_c)}\Big]^2, \\ \mathbb{E}_{q_{\mathrm{obs},t}(\theta_A \mid x,y,\theta_c)} \log \frac{q(y \mid x,\theta_A,\theta_c)}{q(y \mid x,\theta_c)} &= \Delta_{A,t}(x,y,\theta_c), \end{aligned} \tag{36}$$

since the second ratio equals $q_{\mathrm{obs},t}(\theta_A \mid x,y,\theta_c)/q_{\mathrm{obs},t}(\theta_A \mid x,\theta_c)$ by Bayes' rule. The finite factor is stored on the θ_c-axis and retains its channel form; the θ_A-dependent factor, not polynomial in θ_A, is replaced by its posterior expectation, a scalar per entry, as conjugacy requires (Appendix A.3). Because each correction is differentiated under the measure it is defined on, the weights have the form $+D_{\mathrm{KL}} \geq 0$.

Singleton beliefs. ΔF has no singleton dependence, so as in [14] (D.4.6): for the degree-two variables the singleton entropy cancels against the counting term, and for the shared variables only the sum of multipliers is fixed,

$$\begin{aligned} \lambda_{y_1} = -\log \hat{p}_y, \qquad \lambda_{x_0} = -\log p(x_0), \qquad \lambda_{u_1} = -\log p(u_1), \\ \lambda^{\mathrm{obs}}_{x_1} + \lambda^{\mathrm{dyn}}_{x_1} = -\log q^*_{x_1} - 1 - \log \hat{p}_x, \\ \lambda^{\mathrm{obs}}_{\theta_c} + \lambda^{\mathrm{dyn}}_{\theta_c} = -\log q^*_{\theta_c} - 1 - \log p(\theta_c). \end{aligned} \tag{37}$$

Remark 2 (Parameter beliefs and the frozen-parameter update). During planning $q(\theta_A), q(\theta_B)$ are frozen, so m, the one-count posteriors and the weights depend on them only through the fixed concentrations, whereas $q(\theta_c)$ is an ordinary singleton belief and moves within the sweep. Exact functional stationarity of F_T in a free q_{θ_A} has two parts: the conjugate statistic (a Dirichlet count update) and a *centered score* $s_A(\theta_A) = \mathbb{E}_w[\log\beta - \log\bar{\beta} - \Delta_A]$ under the reweighted block measure $w \propto q_{yx,t}\, p(y \mid x,\theta_A,\theta_c)/m_A$, which integrates to zero against q_{θ_A} and vanishes at point mass and as counts grow. The score would reshape q_{θ_A} toward parameter values of above-average posterior distinguishability without shifting count mass. Between environment steps we drop it and update the concentrations by the plain conjugate count update from the realized transition and observation,

$$\begin{aligned} q_{\theta_A} &\propto p(\theta_A)\, \exp\Big(\textstyle\sum_t \mathbb{E}_{q_{yx,t}}[\log p(y_t \mid x_t,\theta_A,\theta_c)]\Big), \\ q_{\theta_B} &\propto p(\theta_B)\, \exp\Big(\textstyle\sum_t \mathbb{E}_{q_{\mathrm{trip},t}}[\log p(x_t \mid x_{t-1},u_t,\theta_B,\theta_c)]\Big), \end{aligned} \tag{38}$$

with $q_{yx,t}, q_{\mathrm{trip},t}$ now the (soft) realized data, the count distributed over contexts according to $q(\theta_c)$, which is itself updated by exact Bayes. This is a partial-gradient step of F_T that preserves exact conjugacy, with the dropped terms exhibited; the score is worth including only if monitoring shows the plain step increasing F_T.

B.3 Solving the Stationarity Equations as Message Passing

Given the kernels (17)–(18) with lagged conditionals r_t, Propositions 3–4 are an exact Bethe belief-propagation problem on the discrete graph over

$\{y_t, x_t, u_t, \theta_c\}$ [18], and the standard factor-to-variable rule applies,

$$\mu_{a\to j}(s_j) \propto \sum_{s_{a\setminus j}} \bar{g}_a(s_a) \prod_{i\in\mathcal{E}(a)\setminus j} \mu_{i\to a}(s_i). \tag{39}$$

Observation factor, writing $\mu_{y_t}, \mu_{x_t}, \mu_{\theta_c}$ for the messages into $f_{\mathrm{obs},t}$ (with $\mu_{y_t} = \hat{p}_y$):

$$\begin{gathered}\mu_{\mathrm{obs}_t\to x_t} = \sum_{y_t,\theta_c} \bar{g}_{\mathrm{obs},t}\, \mu_{y_t}\mu_{\theta_c}, \qquad \mu_{\mathrm{obs}_t\to y_t} = \sum_{x_t,\theta_c} \bar{g}_{\mathrm{obs},t}\, \mu_{x_t}\mu_{\theta_c}, \\ \mu_{\mathrm{obs}_t\to\theta_c} = \sum_{y_t,x_t} \bar{g}_{\mathrm{obs},t}\, \mu_{y_t}\mu_{x_t}.\end{gathered} \tag{40}$$

Dynamics factor: analogously to x_t, x_{t-1}, u_t, θ_c, each summing $\bar{g}_{\mathrm{dyn},t}$ against the other three incoming messages, with $\mu_{u_t\to\mathrm{dyn}_t} = p(u_t)$ and boundary $\mu_{x_0\to\mathrm{dyn}_1} = p(x_0)$. There are no messages to or from $\theta_A^{(t)}, \theta_B^{(t)}$. Block and singleton beliefs are

$$\begin{aligned} q_{yx,t} &\propto \bar{g}_{\mathrm{obs},t}\, \mu_{y_t\to\mathrm{obs}_t}\, \mu_{x_t\to\mathrm{obs}_t}\, \mu_{\theta_c\to\mathrm{obs}_t}, \\ q_{\mathrm{trip},t} &\propto \bar{g}_{\mathrm{dyn},t}\, \mu_{x_t\to\mathrm{dyn}_t}\, \mu_{x_{t-1}\to\mathrm{dyn}_t}\, \mu_{u_t\to\mathrm{dyn}_t}\, \mu_{\theta_c\to\mathrm{dyn}_t}, \end{aligned} \tag{41}$$

$$\begin{gathered} q(x_t) \propto \hat{p}_x\, \mu_{\mathrm{obs}_t\to x_t}\, \mu_{\mathrm{dyn}_t\to x_t}\, \mu_{\mathrm{dyn}_{t+1}\to x_t}, \qquad q(\theta_c) \propto p(\theta_c) \prod_{\tau=1}^{T} \mu_{\mathrm{obs}_\tau\to\theta_c}\, \mu_{\mathrm{dyn}_\tau\to\theta_c}, \\ q(u_t) \propto p(u_t)\, \mu_{\mathrm{dyn}_t\to u_t}, \qquad q(y_t) \propto \hat{p}_y\, \mu_{\mathrm{obs}_t\to y_t}, \end{gathered} \tag{42}$$

which at a fixed point recover (34)–(35) with (37). The full factor beliefs follow from (13), $q_{\mathrm{obs},t} = q_{yx,t}\,\mathrm{Dir}(\alpha_A + e_{y_t x_t \theta_c})$, and are never stored. The block-layer cost is independent of the size of the Dirichlet parameter space and linear in K.

Refresh. After each sweep, from the new block beliefs,

$$\begin{gathered} r_t(y_t \mid x_t, \theta_c) \leftarrow q_{yx,t}(y_t \mid x_t, \theta_c), \qquad r_t(y_t \mid x_t) \leftarrow q_{yx,t}(y_t \mid x_t), \\ r_t(u_t \mid x_{t-1}) \leftarrow q_{\mathrm{trip},t}(u_t \mid x_{t-1}), \qquad r_t(x_t \mid x_{t-1}, u_t) \leftarrow q_{\mathrm{trip},t}(x_t \mid x_{t-1}, u_t), \end{gathered} \tag{43}$$

$$\bar{n}_{x_t\theta_c} \leftarrow r_t(\cdot \mid x_t, \theta_c), \quad \Delta_{A,t}(x_t, y_t, \theta_c) \leftarrow D_{\mathrm{KL}}\big[\mathrm{Dir}(\alpha_A + e_{y_t x_t \theta_c}) \,\big\|\, \mathrm{Dir}(\alpha_A + \bar{n}_{x_t\theta_c})\big], \tag{44}$$

and likewise $\Delta_{B,t}$ with fractional counts $\bar{n}_{x_{t-1}\theta_c}(x_t, u_t) = q_{\mathrm{trip},t}(x_t, u_t \mid x_{t-1}, \theta_c)$ spread over the (x_{t-1}, u_t, θ_c) columns, so that the divergence carries cross-column terms. The kernels are rebuilt immediately. Damping is applied linearly in $\bar{n}$, which keeps the cavity Dirichlet. Since no conditional is built from messages and the one-count posteriors never feed back into themselves, the divergent channel recursion of [14] is structurally absent.

Closed forms. With $q_{\theta_A} = \prod_{x,\theta_c} \mathrm{Dir}(\alpha_A(\cdot \mid x, \theta_c))$: $m_A(y \mid x, \theta_c) = \alpha_A(y \mid x, \theta_c)/\alpha_{A,0}(x, \theta_c)$; $q_{\mathrm{obs},t}(\theta_A \mid x, y, \theta_c) = \mathrm{Dir}(\alpha_A + e_{yx\theta_c})$; the projected cavity is $\mathrm{Dir}(\alpha_A + \bar{n}_{x\theta_c})$; and for two Dirichlets on one column,

$$D_{\mathrm{KL}}[\mathrm{Dir}(a) \,\|\, \mathrm{Dir}(b)] = \log\frac{\Gamma(a_0)}{\Gamma(b_0)} + \sum_k \Big[\log\frac{\Gamma(b_k)}{\Gamma(a_k)} + (a_k - b_k)\big(\psi(a_k) - \psi(a_0)\big)\Big]. \tag{45}$$

The salience $I_{q_{yx,t}}(y_t; \theta_c \mid x_t)$ is a finite sum over the stored array. F_T (16) is assembled from quantities the sweep already computes and gives a stopping criterion on $|\Delta F_T|$.

Schedule. Algorithm 1 of the main text, with the objects of this appendix in place: the predictive likelihoods and the one-count posteriors $\mathrm{Dir}(\alpha_A + e_{yx\theta_c})$, $\mathrm{Dir}(\alpha_B + e_{x'xu\theta_c})$ are built once per planning call from the current counts; each sweep rebuilds the kernels (17)–(18), runs one forward–backward pass of (39)–(42) over the chain, applies the refresh (43)–(44) and evaluates F_T; the policy is read from $q(u_1 \mid x_0)$ at convergence. F_T is a monitoring quantity, not a certificate: block-coordinate descent on a non-convex functional with lagged factors is not guaranteed monotone, and period-two orbits appear as non-monotone F_T, which is the criterion by which damping is selected. Uniform r_t, zero weights and flat parameter beliefs recover plain belief propagation.

B.4 Generic Scheme for Arbitrary T

The corrections (12) are additive in t and every object of Section 4.3 is per-factor, so the family carries one block belief per factor, one shared context belief, and two global, frozen Dirichlet beliefs,

$$\{q_{yx,t},\, q_{\mathrm{trip},t}\}_{t=1}^{T}, \quad \text{singletons incl. } q(\theta_c), \quad q_{\theta_A} = \mathrm{Dir}(\alpha_A),\; q_{\theta_B} = \mathrm{Dir}(\alpha_B). \tag{46}$$

The one-count posteriors and m_A, m_B are time-independent; the cavities, lagged conditionals and weights are time-indexed. Lemmas 4–5 apply per step and give the kernels (17)–(18) at every t. The graph is the temporal chain of [14] with θ_c attached to every factor, so the sweep is loopy belief propagation as there: with state potential $\phi_t(x_t) = \hat{p}_x(x_t)\, \mu_{\mathrm{obs}_t \to x_t}(x_t)$ and context messages $\mu_{\theta_c \to a} = q(\theta_c)/\mu_{a \to \theta_c}$,

$$\overrightarrow{m}_t(x_t) \propto \sum_{x_{t-1}, u_t, \theta_c} \bar{g}_{\mathrm{dyn},t}\, p(u_t)\, \mu_{\theta_c \to \mathrm{dyn}_t}\, \phi_{t-1}(x_{t-1})\, \overrightarrow{m}_{t-1}(x_{t-1}), \tag{47}$$

$$\overleftarrow{m}_t(x_t) \propto \sum_{x_{t+1}, u_{t+1}, \theta_c} \bar{g}_{\mathrm{dyn},t+1}\, p(u_{t+1})\, \mu_{\theta_c \to \mathrm{dyn}_{t+1}}\, \phi_{t+1}(x_{t+1})\, \overleftarrow{m}_{t+1}(x_{t+1}), \tag{48}$$

with $\overrightarrow{m}_0 = p(x_0)$, $\phi_0 \equiv 1$, $\overleftarrow{m}_T \equiv 1$; then $q(x_t) \propto \phi_t\, \overrightarrow{m}_t\, \overleftarrow{m}_t$, the block beliefs follow from (41) with $\mu_{x_t \to \mathrm{obs}_t} = \hat{p}_x \overrightarrow{m}_t \overleftarrow{m}_t$, $\mu_{x_{t-1} \to \mathrm{dyn}_t} = \phi_{t-1} \overrightarrow{m}_{t-1}$, $\mu_{x_t \to \mathrm{dyn}_t} = \phi_t \overleftarrow{m}_t$, and $q(\theta_c)$ from (42). At $T{=}1$ this reduces to (40)–(42). Because every copy is drawn i.i.d. from the same frozen $\mathrm{Dir}(\alpha)$, novelty is scored per factor against the same belief, with no within-horizon diminishing returns; replanning after every real step bounds this to one horizon, and the causal alternative $\alpha_t = \alpha + \sum_{s<t} \bar{n}_s$ (assumed-density filtering along the horizon) retains conjugacy but makes the posteriors and m per-sweep objects. Salience has no such issue: θ_c is shared, so a cue at step s updates $q(\theta_c)$ for every step.

B.5 Limits

All Dirichlet-dependent quantities are available in closed form, so conjugacy is exact and the block-layer cost is independent of the Dirichlet parameter-space size. As $q_{\theta_A}, q_{\theta_B} \to \delta_{\theta^*}$, posterior and cavity coincide, both weights vanish pointwise, $m \to p(\cdot \mid \theta^*, \theta_c)$, and (17)–(18) become the EFE kernels (13a)–(13b) of [13, 14] with their enumerable θ read as θ_c: inference over K known models is the point-mass limit of learning K matrix pairs. The approach is continuous, $D_{\mathrm{KL}}[\mathrm{Dir}(\alpha{+}e) \,\|\, \mathrm{Dir}(\alpha{+}\bar{n})] \to 0$ as counts grow. At $K = 1$ the salience ratio is identically one and the scheme reduces to Dirichlet learning alone, the setting of Section 5. The two epistemic drives differ in scale: novelty is $O(1/\alpha_0)$ and vanishes as counts grow, salience is scale-free and vanishes only when θ_c is resolved; novelty couples no columns, salience couples all of them through the single shared θ_c.

C Experimental Details

C.1 Environments and Protocol

Shared protocol. Four cardinal moves (N, E, S, W; walls and grid edges bump), $K = 1$, one unknown block initialized as a flat Dirichlet with mass 0.1 per entry, the known block clamped by counts of 10^6 times the true matrix. Every planner runs the same exact filter under the posterior-mean matrices m_A, m_B and the count update of Algorithm 1, line 10. Planning: 100 sweeps at damping 0.5, uniform $\hat{p}_y$, horizon $H_s = \min(H, T - s + 1)$ at step s of a T-step episode as in [13], argmax of $q(u_1)$; counts persist across episodes, pose and belief reset, seeds are shared across planners. EFE enumerates the 4^H open-loop policies under m_A, m_B and takes the argmax of $\exp(-G)$, G being the per-policy sum of (7).

Fog-walls maze. 5×5 cells (x, y), y upwards; start $(0, 2)$, goal $(4, 2)$. A is the identity except on the fog block $(1, 1)$, $(1, 2)$, $(1, 3)$, $(2, 2)$, $(3, 2)$, whose columns are uniform over the block, and the noisy cell $(2, 0)$ (mass 0.2 on itself, 0.8 over its neighbours). B is deterministic except at the drift cell $(1, 2)$: intended move w.p. 0.4, north or south w.p. 0.3 each. Cells $(2, 1)$, $(2, 3)$, $(3, 1)$, $(3, 3)$ are walls: the direct 4-step route crosses drift and fog, the 8-step top and bottom rows do not (the latter passes the noisy cell). $T = H = 8$; episodes end at the goal. Learn-A: all 25 columns of A flat, B known; the matrix error is the mean of $|m_A - A|$ over all entries. This is the fog maze of [13] with its four sinks replaced by walls.

Unmapped detour, doors variant. 5×4 grid ($N = 20$), start $(0, 1)$, goal $(4, 2)$, identity A, walls $(1, 2)$, $(2, 2)$, $(3, 2)$. The shortcut $\rightarrow\rightarrow\rightarrow\rightarrow\uparrow$ ends at the ice $(4, 1)$, whence NORTH reaches the goal w.p. $p = 0.5$ and otherwise slips into the absorbing dead end $(4, 0)$. The detour $\uparrow\uparrow\rightarrow\rightarrow\rightarrow\rightarrow\downarrow$ is deterministic through $(0, 2)$, $(0, 3)$, $(1, 3)$, $(2, 3)$, $(3, 3)$, $(4, 3)$; exactly its six (cell, action) columns start

flat, the other actions at those cells (bumps or the step back) are known. $T = 8$, $H = 7$; episodes end at the goal or the dead end, the last transition folded into the counts. *Detour* = goal reached with every visited cell on the detour; err_B = mean $|m_B - B|$ over the six door columns (prior value 0.095). Frozen B keeps the door columns flat and never updates them; the *cells* variant flattens all four actions of the six cells.

C.2 Controls and Robustness

Fog-walls, A, B known (30 seeds). At $H = 8$ every SP-EFE variant puts $q(u_1)$ entirely on NORTH and takes the top row on 30/30 seeds; SP-Pref-Plan ($q(\mathrm{EAST}) = 0.73$) and SP-Pref (0.32) go into the fog and score 18/30 and 17/30: they plan as if they will know where they land, and the planning correction sharpens that commitment. At $H = 11$ bumping after a wrong landing still reaches the goal and the ceiling closes to 29/30 (SP-Pref 22/30). The learn-A policies of SP-Pref(-Plan) equal their known-A policies: their goal rate is the luck of the drift, not a function of m_A.

Unmapped detour. Known B: SP-Pref ($q(\mathrm{EAST}) = 0.30$) takes the ice because the uncorrected Bethe term $-\mathbb{H}[q(u_t \mid x_{t-1})]$ rewards states in which all actions look alike, which absorbing cells do: a sink costs $4 - \log 4 \approx 2.6$ nats per step instead of 4. Frozen B: SP-Pref-Plan gambles ($q(\mathrm{EAST}) = 0.70$, zero door visits); SP-EFE$^{-\mathrm{dyn}}$ farms the first door (0 goals: Δ_B never decays); SP-EFE, SP-EFE$^{-\Delta_B}$ and SP-Pref reach the goal via the detour on 300/300 episodes *without learning*: with frozen flat doors "most uncertain column" and "detour" are the same action at every cell, so any entropy follower is walked along the route († in Table 1).

Robustness. Action indices permuted (W, S, E, N; 200 episodes): SP-EFE$^{-\mathrm{dyn}}$ 198, SP-Pref 101 (25 detours, then the ice), SP-Pref-Plan 89. Mirrored 6×4 layout, 7 doors, $H = 8$ (300 episodes $\times$ 3 seeds): SP-EFE 300, SP-EFE$^{-\mathrm{dyn}}$ 299, SP-Pref 200 ± 6, SP-Pref-Plan 150 ± 5, EFE 100/100. Cells variant (200 episodes $\times$ 5 seeds): SP-EFE$^{-\mathrm{dyn}}$ 159, SP-Pref-Plan 97.6 ± 12.5 with m_B bit-identical to its prior, SP-Pref 185 (off-path flat columns pay its bonus forever), SP-EFE 35.