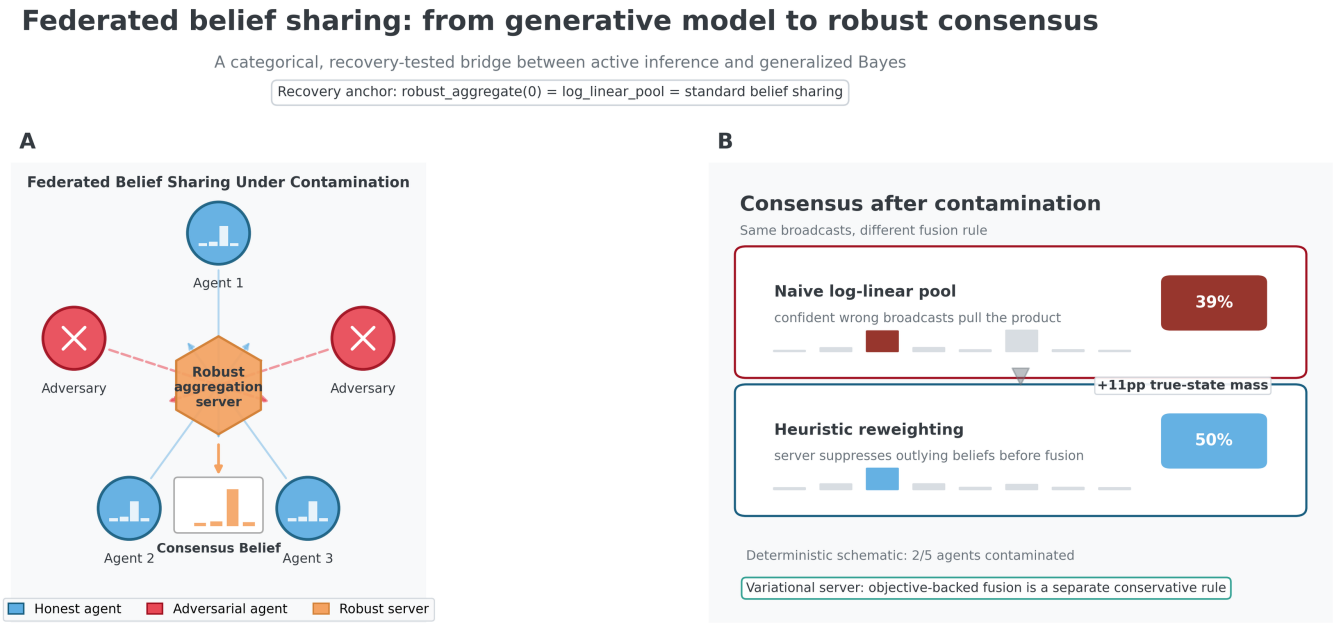
Robust Belief Sharing in Federated Active Inference

A Recovery-Tested Generalized-Variational Framework for Categorical Contamination-Aware Consensus

Daniel Ari Friedman
Active Inference Institute
daniel@activeinference.institute
ORCID: 0000-0001-6232-9096
DOI: 10.5281/zenodo.21864004

August 10, 2026



Contents

1 Abstract

Multi-agent active inference gives a natural account of belief sharing: agents hold local posteriors over a shared latent state, communicate those beliefs, and pool them into a colony-level consensus. The same mechanism is fragile when a member is miscalibrated, corrupted, or strategically wrong. Because the standard pool multiplies the reports together, a single confident-but-wrong broadcast that puts near-zero mass on the true state can pull the whole consensus off it, outweighing many honest members. The colony therefore needs a way to preserve the useful structure of belief sharing while limiting the influence of contaminated beliefs.

This paper presents Active Fedference, a discrete-categorical framework that connects robust federated generalized variational inference with active inference belief sharing. The main bridge is structural: standard belief sharing appears as the non-robust corner of a broader generalized-Bayes family, while robust losses, conservative server fusion, and explicit aggregation diagnostics describe how the system moves away from that corner under declared contamination mechanisms. The result is not a replacement for belief sharing, but a containment result: ordinary belief sharing is recovered when robustness is turned off. Bounded-loss theory applies on the client axis, while the variational-server axis supplies an objective-backed redescending weight update.

The manuscript separates three robustness axes that are often conflated. First, client-side generalized-Bayes updates change how each agent absorbs evidence; this is the rigorous axis, carrying FedGVI's bounded-influence result only under the source theorem's loss, model, and contamination assumptions. Second, a sharp server-side reweighting heuristic suppresses beliefs that pull away from the emerging consensus, while carrying only its recovery-limit guarantee — no proven objective and no bounded-influence bound. Third, a variational aggregation rule supplies a more conservative objective-backed server alternative, with a raw effective-weight bound but not an estimator-level bounded-influence proof for the normalized consensus. Keeping these axes separate lets the paper state exactly which claims are proven, which are empirical, and which remain engineering extensions.

The study suite then exercises the framework as an end-to-end research system: recovery checks anchor the standard-Bayes limit, belief-sharing studies verify the communication baseline, contamination experiments test robust consensus, and extension studies probe moving agents, hierarchical latent structure, sensitivity to acuity and colony size, parameter recovery, and single-host socket-backed federation traces. All reported quantities are generated from deterministic analysis artifacts and injected into the manuscript by token, so the paper, figures, release package, and validation reports remain tied to the same execution record.

The open-source repository is ActiveInferenceInstitute/Active_Fedference. The production Zenodo release DOI is 10.5281/zenodo.21864004, and the repository and deposited PDF point to each other through this DOI and the repository URL.

Keywords: active inference, federated learning, generalised variational inference, belief sharing, robustness, FedGVI

The complete system schematic is shown in fig. 1.

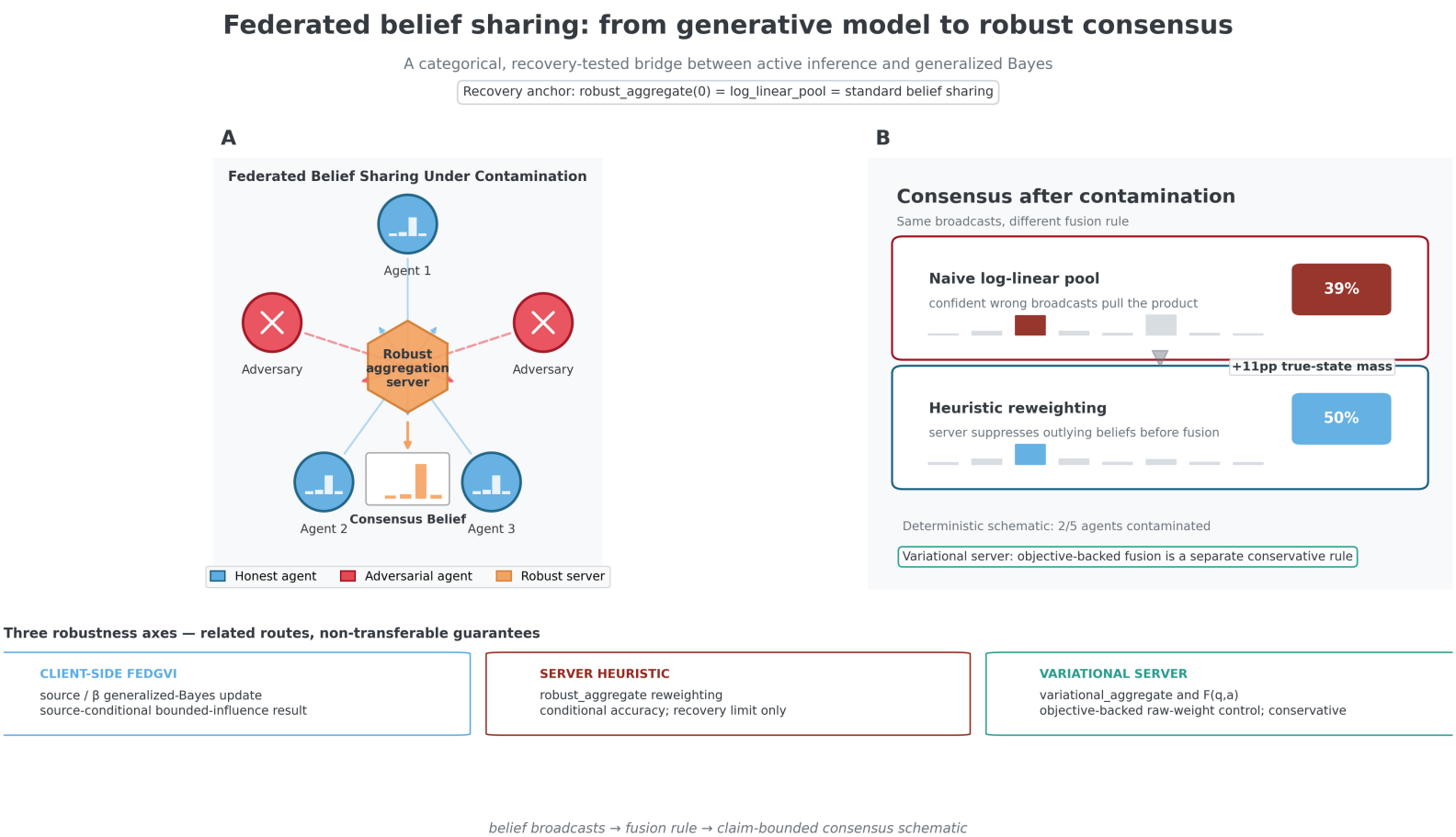


Figure 1: Graphical abstract. Source relation: original project schematic; estimand: component relationships and recovery boundaries; uncertainty: none. **Recovery ribbon:** the zero-robustness identity anchors the construction at the standard log-linear belief-sharing pool. **Network panel:** the federated diagram shows 5 agents (3 honest, 2 adversarial) transmitting categorical beliefs to a central server. x-axis is agent position in the ring layout (left to right); y-axis/rows: each per-agent mini-bar glyph indexes posterior probability mass over hidden states. **Consensus panel:** deterministic outcome cards under 40% adversarial contamination compare the naive pool with canonical `robust_aggregate` heuristic reweighting; the displayed 39% and 50% are computed from the schematic beliefs. **Axis strip:** client-side FedGVI, server-side heuristic, and variational-server claims are shown as separate routes with non-transferable guarantees. This deterministic formal/mechanistic schematic has no CI, error band, or significance marker; it does not assign the variational server's objective-backed property to `robust_aggregate`.

2 Introduction: from belief sharing to robust generalized Bayes

A colony of active-inference agents that shares beliefs inherits both the power and the fragility of the pool it uses: the same multiplication of reports that sharpens an honest consensus lets a single confident, wrong member capture it. Two research communities each hold part of the remedy but do not, on their own, close the gap. The active-inference community gives agents generative models, action selection, and colony-level belief sharing. The robust- and federated-Bayes community gives generalized objectives and aggregation methods for inference under misspecification or contamination. The cited literatures do not, however, provide a single tested treatment of robust categorical belief fusion for active-inference agents. This introduction states the problem and its evidence boundary; sec. 3 scopes the reviewed gap, and sec. 4 lists the contributions together with what each result does not establish.

2.1 Active inference supplies generative agents and shared beliefs

Active inference casts perception, learning, and action as the minimization of variational free energy under a generative model, a program that grew out of the free-energy principle [Friston, 2010] and its process-theory implementation [Friston et al., 2017]. In the discrete state-space setting the formalism is now standardized: an agent carries the categorical tensors A (likelihood), B (transitions), C (preferences), and D_0 (initial priors), infers hidden states by minimizing variational free energy, and selects actions by minimizing *expected* free energy — the sum of a risk term and an ambiguity term that together trade off goal-seeking against uncertainty-resolving behavior [Da Costa et al., 2020]. This synthesis is the substrate we build on. Mature toolboxes make it executable at scale: pymdp provides discrete-state active inference in Python [Heins et al., 2022], and RxInfer delivers reactive message passing for exact Bayesian inference [Bagaev et al., 2023].

The community has also moved from single agents to *collectives*. Surprise-minimizing ensembles reproduce collective animal behavior such as schooling and flocking [Heins et al., 2024]; epistemic communities form when agents share a generative model and exchange evidence [Albarracin et al., 2022]; and collective intelligence has been framed directly in active-inference terms [Kaufmann et al., 2021]. The narrow bridge studied here starts from

posterior broadcasts over one shared categorical factor — a predator's location, say — and forms a project log-linear pool (eq. 6). Under the explicit finite-shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, that weighted geometric pool is a specialization of Friston et al.'s Eq. 7 message-combination term [Friston et al., 2024]. It is not a reconstruction of the source message construction, scheduling, cavity policy, generative factors, or complete protocol. That representation connects the qualified categorical bridge to the classical logarithmic-pooling literature [Genest and Zidek, 1986, Genest et al., 1986], to modern Bayesian treatments of log-pool weights [Carvalho et al., 2023], to product-of-experts geometry in machine learning [Hinton, 2002], and to distributed Bayesian estimators [Tresp, 2000]. In the reproduced baseline, communication lowers mean free energy relative to the matched incommunicado condition (sec. 7.2). Friston et al. [2024] crystallized the colony mechanism into three worked simulations — communicating-colony free-energy convergence, Dirichlet language acquisition, and Bayesian model reduction structure emergence — whose mechanisms motivate three reduced categorical analogues in this repository (sec. 7.1). They are not numerical or exact protocol replications; a source-parity reconstruction remains future work before evaluating source-level equivalence. Alongside fusion, the community has tools for growing the model itself: active inference connects naturally to Bayesian optimal experimental design and model selection [Smith et al., 2022], and post-hoc Bayesian model reduction prunes redundant structure by comparing free-energy bounds [Friston and Penny, 2011] — the engine behind our emergence study (sec. 7.4).

The reviewed active-inference belief-sharing work does not systematically characterize fusion under explicit contamination or intentionally wrong belief broadcasts. The log-linear pool assumes reports are compatible with the shared generative model, while the opinion-pooling literature makes the assumptions about independence, weights, and external Bayesian coherence explicit [Genest and Zidek, 1986, Genest et al., 1986]. Fully Bayesian aggregation sharpens the point: geometric pooling is normatively compelling under dynamic-Bayesian rationality assumptions, not an assumption-free robustness procedure [Dietrich, 2021]. The same product geometry can be brittle: if a report assigns zero or near-zero mass to the true state, the product can assign near-zero mass there too. That is the failure mode tested here, not a claim that every federation or every product pool behaves identically.

2.2 Robust and federated Bayes supplies bounded-influence updating

A separate literature studies bounded-influence updating outside active inference. **Federated learning** aggregates models trained on decentralized data without pooling the data itself, the canonical algorithm being FedAvg [McMahan et al., 2017]. Cast probabilistically, federated and continual learning unify under **partitioned variational inference**, in which each client owns a factor of a global approximate posterior and the server combines factors in natural-parameter space [Bui et al., 2018, Ashman et al., 2022]. This is a closely related factor algebra expressed in a different vocabulary; the implementation tests the correspondence in the categorical recovery limit rather than assuming that the two settings are interchangeable.

The robustness this paper needs comes from **generalized Bayesian inference**, which replaces the likelihood with a *loss* and the KL regularizer with a *general divergence*, so that the posterior minimizes a generalized objective rather than applying Bayes' rule literally [Bissiri et al., 2016, Jewson et al., 2018, Knoblauch et al., 2022]. This includes Gibbs-posterior updates [Jiang and Tanner, 2008], coarsened or tempered posteriors that condition on neighborhoods rather than exact data [Miller and Dunson, 2018], and learning-rate/temperature choices designed for safe updating under misspecification [Grünwald, 2012, Kleijn and van der Vaart, 2012]. Choosing a loss or divergence with a bounded-influence property — for example, the density-power β-loss [Basu et al., 1998, Fujisawa and Eguchi, 2008, Ghosh and Basu, 2015] or generalized cross-entropy [Zhang and Sabuncu, 2018] — can cap the influence of a contaminated observation under the corresponding assumptions. **FedGVI** [Mildner et al., 2025b] federates this idea: each client runs a robust generalized-Bayes update against a *cavity* (the global posterior with the client's own factor removed), and the server aggregates the refreshed factors under a chosen server divergence. The result is federated inference with divergence- and loss-specific robustness guarantees, demonstrated on Bayesian neural networks under label contamination.

Crucially, standard Bayes is a *corner* of this generalized family — the case where the loss is the negative log-likelihood and the divergence is KL. Friston et al. do not claim FedGVI, generalized Bayes, β-divergence, or robustness; this manuscript supplies the recasting. Robust generalized Bayes therefore does not replace exact-Bayes belief fusion; it contains the standard pool as a tested zero-robustness recovery limit. That containment is the hinge this paper tests, and the recovery limits are stated formally as numbered results in sec. 6 and the central identity in sec. 5.8.

The historical framing is deliberately modest. The manuscript uses early probability, inverse-probability, utility, and collective-judgment sources as a conceptual genealogy for belief, evidence, expectation, and aggregation [Pascal and Fermat, 1654, Huygens, 1657, Bernoulli, 1713, de Moivre, 1718, Bayes, 1763, Laplace, 1774, Bernoulli, 1738, Condorcet, 1785]. These sources do not anticipate KL divergence, product-of-experts learning, variational Bayes, or federated optimization; the modern correspondence to FedGVI is the formal construction proved and tested here.

Beyond this core, we evaluate a tempered objective family, a deterministic MLP aggregation transfer, and disjoint-field-of-view communication. These are boundary tests: the first probes the accuracy/weight-control trade-off, the second tests API portability in one additional model class, and the third separates a binary-complement null result from a larger-state-space case where communication materially improves consensus.

2.3 Questions, design, and evidence boundary

The paper answers four scoped questions. First, does turning robustness off recover the standard log-linear pool and closed-form Bayes update? Second, under the declared confident-wrong broadcast mechanism, how does the server heuristic change consensus accuracy across contamination rates? Third, what does the objective-backed variational server rule guarantee, and what accuracy does it trade away? Fourth, how do communication, hierarchy, sensitivity, and parameter recovery behave in the accompanying categorical extensions? The first question is answered algebraically and by machine-precision checks; the others are conditional simulation results. The independent unit, resampling scheme, and fixed hidden-state/attack-target estimand are specified in sec. 5.22 and sec. 5.27.

2.3.1 How to read the visual architecture

The manuscript uses two complementary visual layers. The formal schematics adapt the generative-model and posterior-sharing perspective of Friston et al. [2024] to the categorical implementation: fig. 4 shows the private sensory report and $A/B/C/D$ substrate, fig. 3 shows how three local posterior messages become server inputs, and fig. 5 shows the hidden-state, agent, and active-control context. These diagrams are explanatory

maps, not additional empirical observations. The data-bearing figures then report the executed recovery checks, conditional contamination sweep, and objective/descent diagnostics; their captions identify the relevant uncertainty and resampling unit. The graphical abstract in fig. 1 compresses the same logic into a recovery anchor, a federation pathway, and three explicitly non-transferable robustness axes.

fig. 2 illustrates one configured failure-and-repair case: a partially contaminated colony broadcasts beliefs (Panel A), the equal-weight log-linear pool is pulled toward the attack target (Panel B), and the server-side heuristic reweights the broadcasts (Panel C). It is a deterministic schematic, not a universal robustness claim; the three axes and their distinct guarantees are defined in sec. 3 and sec. 3.3.

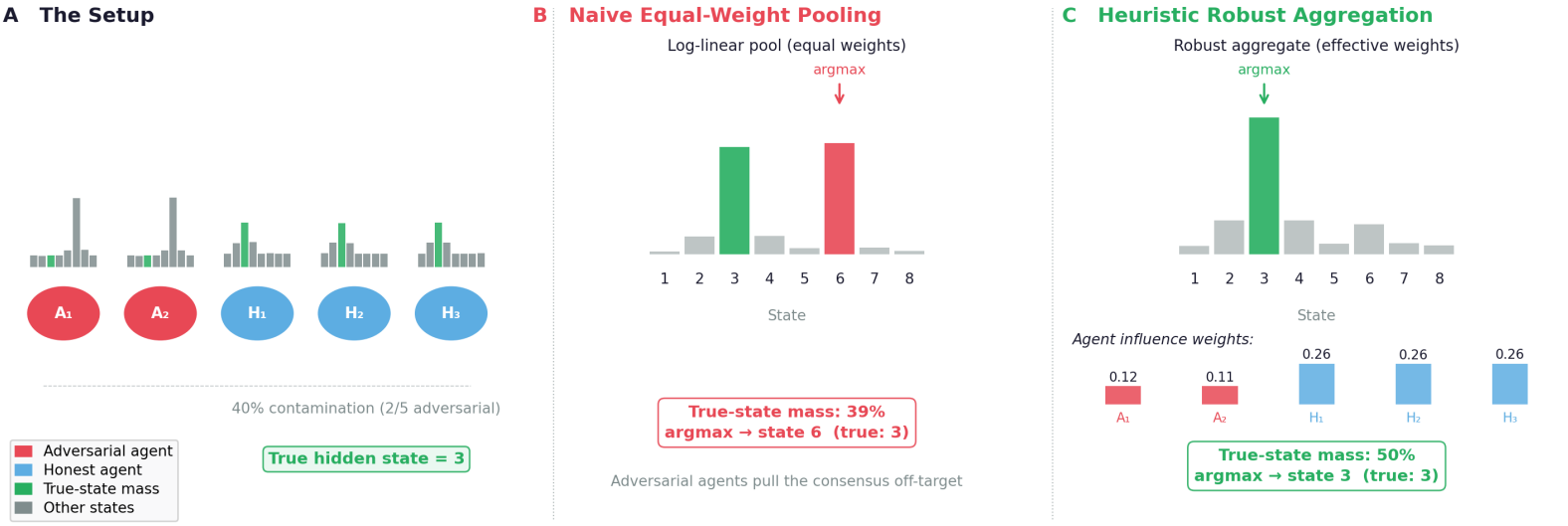


Figure 2: System overview. Source relation: original project schematic; estimand: displayed posterior-mass and influence-weight contrasts; uncertainty: none. 5 agents with heterogeneous beliefs (blue = honest, red = adversarial) feed into naive pooling (Panel B, argmax pulled off-target) versus canonical `robust_aggregate` heuristic reweighting (Panel C, true state recovered). x-axis is hidden-state index (1–8) in Panels B and C; y-axis: bar height indexes probability mass per state, and each agent in Panel A carries its own mini posterior bar chart (green column = true state). The true hidden state is 3; under 40% contamination the equal-weight pool concentrates 39% of consensus mass on the true state (its argmax lands on the adversaries' state), while the heuristic concentrates 50% and recovers the correct argmax. This is a single deterministic schematic (no resampling, hence no error bars or CI); all percentages are computed from the pooled beliefs shown, and the panel does not claim the variational server's objective-backed weight-control result for `robust_aggregate`.

3 Research gap and claim boundary

The two communities of sec. 2 provide substantial pieces of the problem, but the cited threads do not answer the same question. The gap is not a claim that either field is incomplete; it is the untested intersection between active-inference belief sharing and robust generalized Bayes. This section describes that intersection thread by thread, states the scoped bridge evaluated here, and records the evidence boundary that travels with it.

3.1 Five reviewed threads and their open intersection

Five threads in the reviewed literature run toward the gap but do not, in the sources cited here, cross it — three from active inference and two from robust Bayes.

Thread 1 — Generative modeling and action (active inference). Discrete active inference is a mature, synthesized formalism with standardized tensors and an expected-free-energy action rule [Friston, 2010, Da Costa et al., 2020], and it is executable at scale through pymdp [Heins et al., 2022] and RxInfer [Bagaev et al., 2023]. *Boundary:* these sources do not by themselves equip the active-inference inference step with the contamination mechanism evaluated here.

Thread 2 — Collective belief coordination (active inference). Colonies coordinate by sharing beliefs and minimizing collective surprise, reproducing flocking [Heins et al., 2024], epistemic-community formation [Albarracin et al., 2022], and collective intelligence [Kaufmann et al., 2021]. *Boundary:* the cited collective-belief studies treat the broadcast posteriors as trusted and do not evaluate the robustness of fusion to misspecified or intentionally wrong members.

Thread 3 — Belief sharing and structure growth (active inference). Federated belief sharing fuses posteriors into a hive-mind consensus [Friston et al., 2024], and post-hoc Bayesian model reduction [Friston and Penny, 2011] together with optimal-design-flavored model selection [Smith et al., 2022] grow and prune the generative model. *Boundary:* in the cited belief-sharing line, fusion is exact-Bayes and trusting; its connection to the robustness theory developed for federated learning is not evaluated.

Thread 4 — Federated and partitioned inference (robust Bayes). Federated learning aggregates decentralized models [McMahan et al., 2017], and partitioned variational inference gives the factor-algebra framework that unifies federated and continual learning [Bui et al., 2018, Ashman et al., 2022]. *Boundary:* the cited examples target parameter or predictive-model factors rather than the generative-model-bearing, action-selecting POMDP belief consensus used here.

Thread 5 — Generalized and robust Bayes (robust Bayes). Generalized Bayesian updating replaces the likelihood–KL pair with a loss–divergence pair [Bissiri et al., 2016, Knoblauch et al., 2022]; bounded losses — the density-power β-divergence [Basu et al., 1998] and generalized cross-entropy

[Zhang and Sabuncu, 2018] — deliver bounded influence; and FedGVI [Mildner et al., 2025b] federates the robust objective with provable guarantees. *Boundary:* the cited robust-Bayes apparatus does not evaluate active-inference POMDP belief consensus. Its behavior in the discrete categorical regime of the worked belief-sharing example [Friston et al., 2024] is the scoped setting evaluated here.

3.2 The belief-fusion bridge evaluated here

Across these five reviewed threads, the missing intersection is specific: the active-inference sources provide belief fusion but not the contamination analysis used here, while the robust-Bayes sources provide robust inference but not this acting-agent categorical consensus. We evaluate the bridge with robust, generalized-Bayes belief fusion for active-inference ensembles, comprising three components and one recovery anchor. (i) A per-agent, FedGVI-faithful generalized-Bayes update carrying bounded-influence robustness through the β- and rcce-losses (eq. 4, eq. 5). (ii) A complementary server-side divergence-reweighting *heuristic* that discounts each agent by its divergence from the emerging consensus (eq. 7). (iii) A conservative server-side variational rule with a stated aggregation free energy and a redescending raw effective-weight bound (eq. 8). The shared anchor is recovery of the standard-Bayes client limit and the project log-linear-pool server identity in their trusting limits (sec. 6.1). Under the qualified bridge of sec. 5.8, the latter specializes Friston et al.'s Eq. 7 message-combination term [Friston et al., 2024], not the complete source protocol. The algebraic result and machine-precision checks therefore identify a recovery boundary rather than a replacement. Headline comparisons use matched paired statistics — paired Wilcoxon [Wilcoxon, 1945], Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995], bootstrap confidence intervals [Efron and Tibshirani, 1993], and observed-effect design-power planning — and are certified for reproducibility [Peng, 2011].

3.3 Guarantee map: three robustness axes

A red-team review surfaced a distinction we carry through the paper rather than paper over: robustness enters in **three** places, and they do not have the same theoretical standing.

1. **Client-side (rigorous).** The per-agent generalized-Bayes update, driven by a bounded loss inside `generalized_posterior` (eq. 4, eq. 5). It is derived from the stated objective eq. 1 and provably limits to negative-log-likelihood / Bayes — and hence to the standard pool — as the loss parameter goes to zero (Corollary 6 + Proposition 4). This axis inherits FedGVI's [Mildner et al., 2025b] bounded-influence result only under the source theorem's stated loss, model, and contamination assumptions.
2. **Server-side (heuristic).** The divergence-reweighting aggregator `robust_aggregate`, which discounts each agent by $\exp(-c\,\mathrm{KL}(q_n \,\|\, q))$. Only its *recovery* limit is proven — at $c = 0$ it equals the standard log-linear pool (eq. 7, Theorem 5); it is **not** the closed-form minimizer of a FedGVI objective. We present it as a complementary heuristic and never claim it inherits the bounded-influence bound.
3. **Server-side (objective-backed, conservative).** The variational aggregator `variational_aggregate` applies exact block updates that do not increase the stated free energy eq. 8, recovers the same log-linear pool in the trusting limit, and bounds each raw effective weight by its base weight. Its honest cost is conservatism: it is not the sharp accuracy-maximizing heuristic.

The robustness sweep (sec. 7.5) reports these axes, and sec. 8.15 states which claim rests on which. No figure, table, or sentence in this paper grants the server-side heuristic the guarantee that belongs to the client-side update or the variational server objective: the effect-size, confidence-interval, and power enrichment that decorate the sweep characterize the heuristic's *behavior*, not a per-agent or variational guarantee. This honesty is the point: the client-side axis is source-theorem-backed, the sharp server heuristic is useful but clearly labeled, and the variational server axis is rigorous but conservative.

4 Contributions and evidence boundaries

We make ten contributions, each paired with a theorem, figure, table, or generated token and with an explicit boundary on what the evidence establishes. They fall into four groups: the first two build the core and its recovery contract; the next two report and statistically qualify the contaminated-consensus result; the fifth and sixth make the robustness axes explicit and add the objective-backed server rule; and the remaining four are scoped extensions — parameter recovery, the tempered aggregation family, an aggregation-API transfer, and a disjoint-observation communication test.

1. **A discrete-categorical FedGVI core.** A typed, deterministic, pure-NumPy/ SciPy reimplementation of the FedGVI [Mildner et al., 2025b] generalized-Bayes primitives — divergences, bounded robust losses, the generalized posterior, the cavity/factor algebra, and robust aggregation — in the discrete-categorical setting that active inference [Da Costa et al., 2020] uses. The objective (eq. 1) and its closed-form tempered-softmax solution are stated in Definition 1 and tested by the recovery-limit probes. The whole core is zero-mock and reproducible [Peng, 2011].
2. **A recovery-tested connection between categorical pooling and robust Bayes.** A recovery certificate showing that the client KL/negative-log-likelihood loss limits recover Bayes and the server's zero-robustness branch recovers the project log-linear pool. Under the explicit shared-support, posterior-log-potential, and fixed-weight assumptions in sec. 5.8, that pool specializes Friston et al.'s Eq. 7 message-combination term [Friston et al., 2024], not the complete source protocol. Three recovery limits are pinned to bit-level residuals: the bounded losses recover NLL/Bayes (Corollary 6 + Proposition 4; ≤ 0 and ≤ 0 maximum residual, eq. 6), the Rényi divergence recovers KL (Lemma 3; ≤ 0 residual, eq. 3), and the server-side reweighting pool recovers the naive pool in its trusting limit (Theorem 5; ≤ 0 residual, eq. 7). Robustness is thereby a tested recovery-limit extension, not a replacement.
3. **End-to-end evaluation of robust federated active inference.** Three worked categorical source-mechanism analogues — communicating colonies reaching lower free energy (sec. 7.2), Dirichlet language acquisition (sec. 7.3), and structure emergence by Bayesian model reduction [Friston and Penny, 2011] (sec. 7.4) — plus a contaminated-sentinel robustness sweep (sec. 7.5) in which the naive pool degrades and at least one server-side robust member clears the configured threshold at the most severe swept rate.
4. **A statistically qualified server-side contrast.** The "robust beats naive" conclusion for the declared contamination rate is produced only by a matched-pairs Wilcoxon signed-rank test [Wilcoxon, 1945] deflated across the divergence family with Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995], reported with bootstrap confidence intervals [Efron and Tibshirani, 1993] and observed-effect design-power planning.

Across 960 paired trials the headline display method (RKL; tied set: RKL, AR, beta, rcce) reaches accuracy 0.9867 against the naive pool's 0.9021 at the verdict rate ($q = 1.11 \times 10^{-158}$, rank-biserial-derived d-equivalent = saturated (r=+1)). The predeclared selection rule is largest positive rank-biserial effect_size; stable method order tie-break; the method with the largest paired mean difference is AR. Every headline number is a generated token.

5. **An explicit accounting of the robustness axes.** A clear separation (sec. 3.3) between the client-side per-agent update, which inherits FedGVI's bounded-influence result under its stated source assumptions; the sharp server-side reweighting heuristic, whose positive formal property is its recovery limit and whose declared separable objective class has a scoped no-go result; and the conservative variational server rule, which is objective-backed and has a redescending effective-weight update but is not the accuracy-maximizer. Downstream users are told exactly which result is theoretically backed and which is a labeled heuristic.
6. **An objective-backed server aggregator with redescending weights.** We derive an aggregation free energy eq. 8 whose exact block updates define the `variational_aggregate` rule (sec. 5.8.2, sec. 11): each exact block update monotonically decreases a stated objective, a converged fixed point is coordinatewise stationary, and the implementation keeps the lowest observed objective among converged configured starts (or reports the best unfinished trace as non-converged), recovers the standard log-linear pool in the trusting limit (eq. 7), and — unlike the sharp heuristic — carries a proven raw effective-weight bound (fig. 14, fig. 15). The honest cost, stated plainly, is conservatism: it is a maximum-entropy-biased consensus and trades peak point-accuracy for that control, so it complements rather than replaces the sharp heuristic of contribution 5.
7. **Executed finite-grid acuity-recovery experiment.** At each value in the 0.60, 0.70, 0.80, 0.90 acuity grid, the study generates 200 synthetic observations in each of 960 trials and selects acuity by marginal-likelihood grid search over the declared finite grid. The observed mean absolute error is 0.0232 with $R^2 = 0.9999$ (fig. 28). Acuity-by-colony-size behavior belongs to the separate sensitivity study; it is not a parameter-recovery result.
8. **The $\mathbf{F}_\lambda$ tempered aggregation family.** A one-parameter $\lambda > 0$ generalization of the variational aggregate (sec. 11.5): $F_\lambda(q, a) = \sum_n a_n \cdot \mathrm{CE}(q, q_n) - \lambda H(q) + (1/c)\,\mathrm{KL}_{\mathrm{gen}}(a \| w)$. At $\lambda = 1.0$ the temperature is unity and the objective reduces to the standard variational aggregate bit-for-bit. Lower λ sharpens the variational q-block toward a maximizing state; it does not algebraically recover `robust_aggregate`. The raw effective-weight update and its bound are preserved for all λ. Full derivation in sec. 11.5.
9. **An aggregation-API transfer demonstration.** The same `robust_aggregate` API that governs the POMDP studies is exercised unchanged with one deterministic MLP trained with the density-power β-loss (16 hidden units, $\beta = 0.5$; generalized variational inference with a point-mass variational family) as the per-client model, supporting portability of the server API to this additional model class (sec. 7.6) when the optional `torch` extra is installed (sec. 5.26); without it the MLP run is skipped and its tokens render accordingly.
10. **Communication benefit under disjoint observations.** A multi-agent extension (sec. 13.7.1) in which 3 agents each observe a 2-slot disjoint window shows that belief sharing materially improves over isolated-agent accuracy in the declared configuration — isolated agents clear the 0.167 chance baseline but stay far below the communicating consensus, which itself remains well short of full accuracy: across 128 seeds isolated accuracy is 0.326 versus communicating 0.493, a reproducible margin under the declared matched-seed comparison (Wilcoxon $p = 0.0000$); this is evidence for the configured disjoint-observation protocol, not a universal communication theorem.

The remainder of the paper proceeds as follows. sec. 5 develops the FedGVI core and the recovery limits; sec. 6 states the numbered recovery theorems and the expected-free-energy identity; sec. 5.22 fixes the configuration; sec. 7 reports the 9 studies, beginning with the recovery checks (sec. 7.1); sec. 8 and sec. 9 synthesize; and sec. 10 and sec. 8.15 document determinism, scope, limitations, and the standing of each robustness axis.

5 Methods: the federated active-inference stack

This section develops the federated generalized variational inference (FedGVI) core in the discrete-categorical setting and defines the primitives whose recovery limits sec. 6 then states as numbered theorems. Every belief here is a categorical pmf — a non-negative vector summing to one — so the generalized-variational-inference machinery reduces to closed forms that are exactly testable. All mathematics lives in `src/fedference/`; the prose names the module and the identity that pins each claim.

The active-inference community has built a rich apparatus for federated belief sharing: discrete-state-space agents that broadcast posteriors and fuse them into a consensus [Da Costa et al., 2020, Friston et al., 2024], message-passing toolboxes that make exact-Bayes inference scalable [Heins et al., 2022, Bagaev et al., 2023], and collective and multi-agent formulations in which ensembles coordinate by sharing observations and beliefs [Heins et al., 2024, Albarracin et al., 2022, Kaufmann et al., 2021]. We accept that apparatus and extend it: the consensus rule the field uses is exact-Bayes and trusting, with no account of what happens when an agent in the ensemble is misspecified or adversarial. Outside active inference, the federated-learning and robust-Bayes literatures address important parts of that question — decentralized aggregation [McMahan et al., 2017], partitioned and federated variational inference [Ashman et al., 2022, Bui et al., 2018], and generalized, robustness-bearing Bayesian updating [Bissiri et al., 2016, Knoblauch et al., 2022, Basu et al., 1998, Zhang and Sabuncu, 2018] — but none has been carried into the generative-model-bearing, action-selecting POMDP setting. The methodology below is the bridge: it federates the FedGVI objective [Mildner et al., 2025b] per agent inside an active-inference ensemble, proves the standard-Bayes client limits, and tests the project-local zero-robustness log-linear-pool identity. Under the qualified categorical bridge of sec. 5.8, that pool specializes Eq. 7's message-combination term rather than the complete source protocol. fig. 2 illustrates the three-axis architecture and the recovery hierarchy.

5.1 Federation protocol: local update, server fusion, broadcast

A colony of N agents shares a single latent factor $s \in \{1, \ldots, n_s\}$ (in the sentinel scenario, the location of a creature on a grid of n_s cells). Each round proceeds in three steps:

1. **Local inference.** Agent n observes o_n and forms a local posterior $q_n(s)$ over the shared factor by a generalized-Bayes update against its own cavity (the colony belief with agent n's previous contribution removed). This is where robustness enters per agent: the update minimizes a

loss-plus- divergence objective, and the FedGVI choice of a bounded loss is what carries the source theorem's bounded-influence result under its matching assumptions.

2. **Broadcast.** Agent n broadcasts $q_n(s)$, optionally with a scalar base weight $w_n \geq 0$.
3. **Aggregation.** The server (or, equivalently, each agent acting as its own server) fuses the broadcast beliefs into a consensus. Following sensory attenuation — "agents do not hear themselves" — an agent's heard consensus excludes its own message.

The protocol has two distinct places where robustness can live, and we keep them separate throughout. The **per-agent generalized-Bayes update** in step 1 is FedGVI-faithful at the stated primitive level: its formal bounded-influence claim is conditional on the source theorem's assumptions. The **server-side aggregation rule** in step 3 admits an optional divergence-reweighting heuristic that down-weights agents far from the emerging consensus; this heuristic is a complementary device whose positive formal property is recovery of the naive consensus in its trusting limit, while a scoped proposition rejects one declared separable objective class. sec. 3.3 holds this boundary; no figure, table, or sentence in this work grants the server-side heuristic the per-agent FedGVI guarantee.

5.2 Notation for beliefs, losses, and divergences

The authoritative symbol and API contract is sec. 13. In the main text, $q_n(s)$ denotes agent n's local posterior, $q(s)$ the global consensus, and $q_{-n}(s)$ the cavity after removing the site factor $t_n(s)$. The prior is $\pi_0(s)$, while π is a policy. The POMDP tensors are $A[o,s]$, $B[s',s,u]$, $C[o]$, and $D_0[s]$. The aggregation weights are w_n (raw/base), a_n (raw variational effective), and $\tilde{a}_n$ (normalized influence). The server robustness coefficient is c, the variational entropy weight is λ, the Rényi order is α, the density-power parameter is β, and the robust cross-entropy parameter is q_{loss}. The notation supplement also defines the seed/trial nesting and all statistical quantities used below.

The study is run over a fixed ensemble of 7 agents sharing a factor of 9 locations, with all randomness seeded at 0; the full per-study configuration is tabulated in tbl. 1. As an independent generative-model-free baseline, we also implement FedGVI in a deterministic MLP complement trained with the density-power β-loss — generalized variational inference with a point-mass variational family (sec. 7.6). The remaining methodology subsections develop each primitive in turn: the generalized-Bayes update and its recovery to standard Bayes (sec. 5.3), the divergence family and its KL limit (sec. 5.6) and the robust loss family and its NLL limit (sec. 5.7), the aggregation identity (sec. 5.8), the lift to a belief-sharing round (sec. 5.9), and the paired statistics that earn every "robust beats naive" verdict (sec. 5.27).

5.3 Generalized Bayes: the route back to standard Bayes

The inference engine FedGVI federates is the generalized (Gibbs) posterior [Bissiri et al., 2016, Jiang and Tanner, 2008, Jewson et al., 2018, Knoblauch et al., 2022], which trades the likelihood for a loss L and the KL regularizer for a general divergence $\mathcal{D}$:

$$q_n^*(s) \;=\; \arg\min_{q_n} \; \mathbb{E}_{q_n}\Big[\textstyle\sum_i L(s;o_i)\Big] \;+\; \tfrac{1}{\tau}\,\mathcal{D}(q_n \,\|\, \pi_0), \tag{1}$$

with prior π_0, learning rate τ, and regularizing divergence $\mathcal{D}$. The learning rate is part of the inferential specification, not a cosmetic constant; coarsened-posterior and safe-Bayes work show why calibration of that temperature matters under misspecification [Miller and Dunson, 2018, Grünwald, 2012], where ordinary Bayes concentrates around a KL pseudo-truth rather than literal truth when the model family is wrong [Kleijn and van der Vaart, 2012]. We name the object eq. 1 defines.

Definition 1 (Generalized-(Gibbs)-Bayes posterior)**.** For a loss L, prior pi_0, learning rate $tau > 0$, and divergence $\mathcal{D}$, the generalized-Bayes posterior is the minimizer q_n^* of (1). For $\mathcal{D} = \mathrm{KL}$ the minimizer is the tempered softmax

$$q_n^*(s) \;\propto\; pi_0(s)\,\exp(\,-\,tau \textstyle\sum_i L(s;o_i)),$$

implemented in `generalized_bayes.generalized_posterior`.

The tempered softmax of eq. 1, stated in the definition above, is not an approximation: it is the exact closed-form minimizer of eq. 1 when the regularizer is the KL divergence, because the categorical support is finite and the objective is strictly convex in q. The recovery to standard Bayes follows by choosing the loss. With $L = \mathrm{NLL}$, $\mathrm{NLL}(p,o) = -\log p(o)$, the exponential in that tempered softmax becomes a product of likelihoods and the minimizer is *exactly* standard Bayes; eq. 6 in sec. 5.8 states that corner, and Corollary 6 there pins it to the closed-form prior-times-likelihood product. The largest observed discrepancy between `generalized_posterior` in this regime and the analytic Bayes posterior is 5.55e-17, reported in sec. 7.1 — exact to machine precision (a maximum deviation of about one ULP), not merely close.

FedGVI computes each client update against a *cavity* rather than the full posterior, so a contributing agent does not double-count its own previous message. We name that operation.

Definition 2 (Cavity / PVI factor update)**.** The cavity removes agent n's factor from the colony posterior in natural-parameter (log) space,

$$q_{-n}(s) \;=\; \frac{q(s)/t_n(s)}{\sum_{s'} q(s')/t_n(s')} \;=\; \mathrm{softmax}(\log q(s) - \log t_n(s)),$$

where the final expression makes the normalization explicit; the partitioned-variational-inference (PVI) update re-multiplies a refreshed factor onto the cavity of (2). Taking a cavity and re-multiplying the original site factor restores the global posterior

$$q(s) = \frac{q_{-n}(s)t_n(s)}{\sum_{s'} q_{-n}(s')t_n(s')}, \tag{2}$$

with the original site factor, the recombination identity, the property `generalized_bayes.cavity` and `generalized_bayes.update_factor` satisfy.

The numbered recombination identity is eq. 2.

The cavity of eq. 2 is the discrete analogue of the expectation- propagation / partitioned-VI cavity used outside active inference [Ashman et al., 2022, Bui et al., 2018], imported here so that the per-agent generalized-Bayes update of eq. 1 is computed against the colony belief with the agent's own contribution removed — exactly the sensory-attenuation discipline the belief-sharing round of sec. 5.9 requires. What remains unspecified in eq. 1 are its two ingredients — the divergence $\mathcal{D}$ and the loss L — whose robust members and standard-Bayes limits sec. 5.6 develops next; the aggregation identity (sec. 5.8) then federates the resulting per-agent posteriors.

The authoritative notation supplement makes the same normalization and recombination contract explicit in eq. 30 and eq. 31; those equations govern the symbols used by the implementation and all later supplements.

5.4 Conjugate likelihood learning for the shared model

Active-inference agents learn the parameters of their generative model, not just plan with them [Smith et al., 2022, Friston et al., 2024]. The likelihood matrix A carries a Dirichlet prior with concentration a over each column, updated conjugately by accumulating observation-state co-occurrence counts (eq. 14), giving the column-normalized expected likelihood. The update of eq. 14 is driven by the expected sufficient statistics under the data-generating model, so as the concentrations accumulate $\mathbb{E}[A]$ converges to the true likelihood. Convergence is measured by the per-column KL divergence summed over hidden states, which decreases monotonically toward the standard-Bayes fixed point; sec. 7.3 reports the learning curve, where the KL falls from 3.4231 to 0.0027 across 24 count batches. A forgetting hyperprior optionally decays the running mass toward an asymptote so the agent does not become infinitely confident; with the hyperprior disabled the classical unbounded accumulation of eq. 14 is recovered. The implementation is `dirichlet_learning.learn_likelihood`.

5.5 Bayesian model reduction for structure comparison

Structure learning in the active-inference frame proceeds by Bayesian model reduction (BMR): given a full model with Dirichlet posterior `post` under prior `prior`, the change in negative variational free energy from swapping in a *reduced* prior — for example one that prunes a redundant column toward zero — is available in closed form without re-running inference [Friston and Penny, 2011, Smith et al., 2022]. Because the likelihood is shared, the reduced posterior is `post + reduced_prior - prior`, and the free-energy difference is a difference of log multivariate Beta functions (eq. 16), where $\ln B(a) = \sum_k \ln\Gamma(a_k) - \ln\Gamma(\sum_k a_k)$ is the log Dirichlet normalizer. A positive ΔF in eq. 16 means the reduced model has more evidence — the pruned structure was redundant and should be adopted; a negative ΔF means the reduction destroyed something the data support. When the reduced prior equals the prior the score is identically zero, the no-reduction fixed point. sec. 7.4 reports $\Delta F = 3.68$ for a redundant reduction (accepted) against $\Delta F = -27.67$ for a supported one (rejected). The implementation is `bayesian_model_reduction.reduce`.

5.6 Divergences: robust objectives and the KL limit

The generalized-Bayes objective eq. 1 has exactly two tunable ingredients: the divergence D that regularizes the update toward the prior or cavity, and the loss L that measures data fidelity. This section develops both — the divergence family first, the robust loss family in sec. 5.7 — and shows that each carries a limit in which it collapses to its standard counterpart, KL for the divergence and NLL for the loss. Those client-side limits establish recovery to standard Bayes. The distinct categorical server bridge in sec. 5.8 then identifies a qualified log-linear message-combination specialization; it does not recover the complete source belief-sharing protocol.

The regularizing divergence D decides how far a client's updated belief may move from its cavity, so choosing D is a modeling decision rather than a numerical detail. The family lives in `divergences.py`. We implement the forward KL (the standard-Bayes case), the reverse KL (FedGVI's `RKL` client divergence), the standard α-Rényi diagnostic, FedGVI's Alpha-Rényi normalization (AR), and total variation (a bounded distance in $[0, 1]$). The single most important recovery property is that the robust members recover the KL divergence in a limit:

$$D_\alpha(q \,\|\, p) \xrightarrow[\alpha\to 1]{} \mathrm{KL}(q \,\|\, p). \tag{3}$$

Lemma 3 (KL is the $\alpha \to 1$ limit of the Rényi family)**.** *For categorical pmfs q, p on a finite support, the α-Rényi divergence $D_\alpha(q \,\|\, p) = (\alpha - 1)^{-1} \log \sum_k q_k^\alpha p_k^{1-\alpha}$ tends to $\mathrm{KL}(q \,\|\, p)$ as $\alpha \to 1$, the limit (3). The `divergences.py` implementation switches to the KL closed form inside a small band around $\alpha = 1$, so on that band the equality is exact rather than merely asymptotic.*

KL is the divergence that makes generalized Bayes collapse to standard Bayes. When local posteriors are then combined by the separately specified categorical message-combination specialization in sec. 5.8, the project recovers its log-linear-pool corner; neither step reconstructs the complete belief-sharing protocol of Friston et al. [Friston et al., 2024]. Everything robust is a controlled departure from that fixed point; Lemma 3 is the formal hinge, and the largest observed Rényi-versus-KL discrepancy in the recovery band is 0 (reported in sec. 7.1).

The standard Rényi diagnostic is `renyi_divergence`; FedGVI's `AR` regularizer is `alpha_renyi_divergence`, equal to the standard form divided by α. For the finite categorical support, `generalized_posterior` solves the named Alpha-Rényi objective through its scalar normalization condition rather than using a generic power-softmax shortcut. This distinction keeps the reported limit and the implemented objective aligned. The `AR` regularizer is not merely a diagnostic: it is exercised as a client divergence in the categorical FedGVI baseline of sec. 7.6, where it pairs with the rcce loss of sec. 5.7 to constitute the genuine per-client robustness axis.

5.7 Robust losses: bounded influence at the Bayes corner

The data-fidelity term of eq. 1 lives in `losses.py`. Standard Bayes uses the negative log-likelihood, $\mathrm{NLL}(p, o) = -\log p(o)$, which is *unbounded*: a single contaminated observation with $p(o) \to 0$ dominates the posterior. This is precisely the fragility the robust-Bayes literature was built to remove [Basu et al., 1998, Fujisawa and Eguchi, 2008, Ghosh and Basu, 2015, Zhang and Sabuncu, 2018], extended into robust-divergence variational inference [Futami et al., 2018], and the property FedGVI imports into federated inference [Mildner et al., 2025b]. The robust-statistics vocabulary

here is the usual influence-function one [Huber and Ronchetti, 2009]: bounded losses reduce the leverage of extreme observations, while NLL does not. We implement two categorical robust losses, each of which recovers NLL in a limit.

The density-power (β) loss [Basu et al., 1998, Fujisawa and Eguchi, 2008, Ghosh and Basu, 2015, Futami et al., 2018] is recentered so that the scalar limit is exact:

$$L_\beta(p, o) \;=\; -\frac{p(o)^\beta - 1}{\beta} \;+\; \frac{\sum_k p_k^{\beta+1} - 1}{\beta + 1}, \qquad L_\beta \xrightarrow[\beta \to 0]{} \text{NLL}. \tag{4}$$

The robust categorical cross-entropy (generalized cross-entropy) [Zhang and Sabuncu, 2018] is

$$L_{q_{\text{loss}}}(p, o) \;=\; \frac{1 - p(o)^{q_{\text{loss}}}}{q_{\text{loss}}}, \qquad L_{q_{\text{loss}}} \xrightarrow[q_{\text{loss}} \to 0]{} \text{NLL}, \tag{5}$$

which by l'Hôpital recovers NLL as $q_{\text{loss}} \to 0$ and at $q_{\text{loss}} = 1$ is the bounded mean-absolute-error loss $1 - p(o)$, finite exactly where NLL diverges.

Proposition 4 (β-loss and rcce recover NLL)**.** *The recentered density-power loss L_β of (4) tends to the negative log-likelihood as $\beta \to 0$, and the robust categorical cross-entropy $L_{q_{\text{loss}}}$ of (5) tends to the negative log-likelihood as $q_{\text{loss}} \to 0$. Both limits are exact in the implementation; the largest observed $\beta \to 0$ discrepancy from the NLL closed form is 0 and the largest $q_{\text{loss}} \to 0$ discrepancy is 0 (Section 7.1). At the bounded end the loss stays finite where NLL diverges, the source of the robustness validated in Section 7.5.*

Taking the loss-parameter limits ($\beta \to 0$ or $q_{\text{loss}} \to 0$) reproduces standard Bayes. Combining those local posteriors through the qualified categorical specialization in sec. 5.8 is a separate server step, not a recovery claim for the complete belief-sharing protocol of Friston et al. [2024]. This is the **per-agent rigorous robustness axis**: it is derived from eq. 1 and provably limits to Bayes through Proposition 4 and Lemma 3, and it is the axis that carries FedGVI's bounded-influence guarantee under the cited matching assumptions. The complementary server-side divergence-reweighting heuristic of sec. 5.8 is a distinct device and is never granted this guarantee (sec. 3.3).

5.8 Aggregation and message passing: standard pool, heuristic, and variational server

The server step lives in `aggregation.py`, where a categorical specialization of the active-inference belief-sharing relation [Friston et al., 2024] and the FedGVI objective [Mildner et al., 2025b] meet. Each agent n broadcasts a categorical local posterior $q_n(s)$ over the shared latent factor, optionally with a scalar base weight w_n. Two fusion rules act directly on these broadcasts, and a third — the objective-backed `variational_aggregate` of sec. 5.8.2 — refines the second into descent on a stated objective. The first is the **log-linear pool**, a project-local product-of-experts consensus. In the terminology of opinion pooling it is the logarithmic pool, a weighted geometric aggregation rule whose Bayesian-coherence assumptions have been studied independently of active inference [Genest and Zidek, 1986, Genest et al., 1986, Carvalho et al., 2023]; in machine-learning terms it is the product-of-experts normalization of local posteriors [Hinton, 2002]:

$$\text{log\_linear\_pool}(\{q_n\}) \;=\; \text{softmax}\Big(\textstyle\sum_n w_n \log q_n\Big), \tag{6}$$

For the source bridge, fix one finite shared support $\mathcal{S}$ with $q_n(s) > 0$ for every agent and state. Suppose the inputs to Eq. 7's softmax message-combination term can be represented as posterior log potentials $m_n(s) = \log q_n(s) + \kappa_n$, where κ_n is constant in s, and use declared fixed weights w_n that do not depend on the emerging consensus (the unweighted case sets each $w_n = 1$). Additive constants then cancel under softmax, giving exactly eq. 6. This is a categorical posterior-log-potential specialization of the source equation's message-combination term, not a reconstruction of source message construction, self-exclusion/cavity policy, scheduling, generative factors, or the complete protocol. The code alias `friston_belief_share` names this qualified specialization only.

The second rule is **`robust_aggregate`**, an iteratively-reweighted pool that discounts each agent by $\exp(-c\,\mathrm{KL}(q_n \,\|\, q))$ against the emerging consensus q. A confidently-wrong (contaminated) agent sits far from the consensus, can earn a small effective weight and be suppressed in the declared diagnostic regimes. This independently motivated rule does not transfer FedGVI's client theorem to the server side: it is the **heuristic robustness axis** of sec. 3.3, distinct from the per-agent rigorous axis of sec. 5.7. It is also only an analogy to robust federated aggregation methods such as divergence-weighted gamma-mean aggregation, geometric-median robust aggregation, or Byzantine-tolerant gradient aggregation [Li et al., 2022, Pillutla et al., 2022, Blanchard et al., 2017]: those methods motivate the risk surface, but they do not supply this rule's guarantee.

The defining identity is bit-level: at zero robustness the reweighted pool is the log-linear pool unchanged.

$$\text{robust\_aggregate}(0) \;\equiv\; \text{log\_linear\_pool}\,. \tag{7}$$

This is an exact project-local code identity. Under the stated posterior-log-potential assumptions, its right-hand side specializes the message-combination term of Eq. 7; the identity itself neither recovers nor certifies the complete source protocol [Friston et al., 2024].

5.8.1 Protocol map: local updates, broadcast, and server fusion

The visual map in fig. 3 makes the protocol boundary explicit: each client updates and broadcasts a categorical posterior; the server chooses the standard pool, heuristic, or variational route. This is a mechanistic schematic, not an additional benchmark: client-side FedGVI is source- conditional, server-heuristic accuracy is conditional on declared contamination, and the variational route owns objective/descent/raw-weight properties.

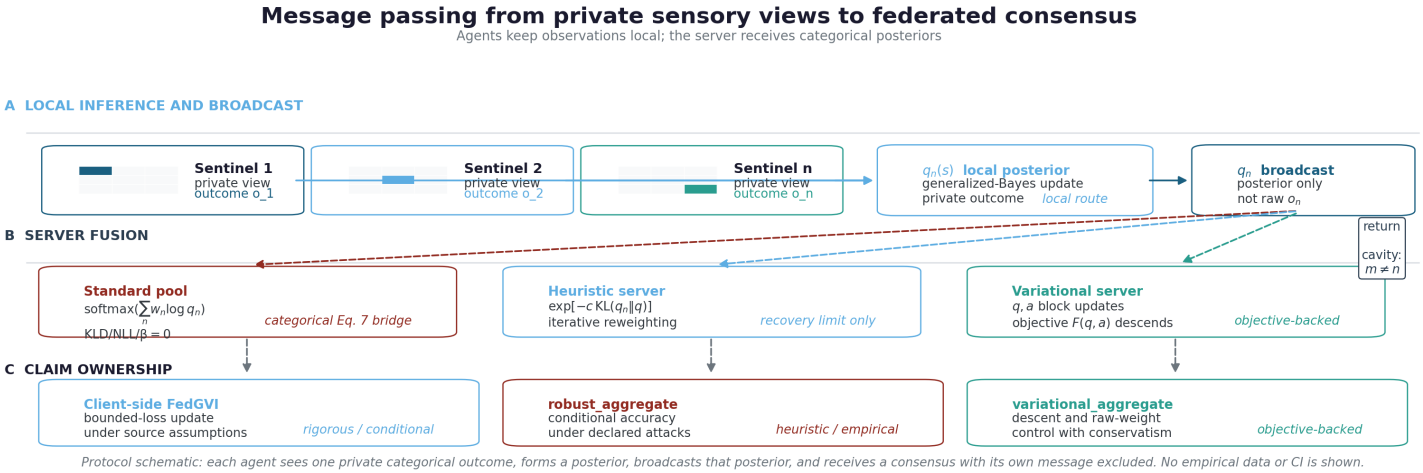


Figure 3: Message-passing schematic for Active Fedference. Source relation: source-inspired original schematic related to Friston et al. (2024), Eq. 7 and Fig. 5; estimand: protocol stages and claim ownership; uncertainty: none. The x-axis is protocol stage from private outcome through local update, posterior broadcast, server fusion, and return; the y-axis uses lanes for local inference, server fusion, and claim ownership. Panel A shows three sentinel agents beginning with private categorical views over the nine-cell location space, converting those views into local posteriors, and broadcasting posteriors rather than raw outcomes. Panel B shows the same broadcast entering the standard log-linear pool, the server-side robust_aggregate heuristic, or the objective-backed variational_aggregate; Panel C keeps their claim ownership separate. The standard pool combines the client KL/NLL/$\beta = 0$ recovery with the qualified categorical Eq. 7 message-combination specialization, while the heuristic retains recovery-limit status only. The return annotation marks cavity exclusion: an agent does not hear its own message. This deterministic formal/mechanistic schematic contains no empirical curve, error band, or confidence interval.

5.8.1.1 Visual protocol map (schematic)

Theorem 5 (Categorical message-combination specialization and local recovery)**.** *Let $\mathcal{S}$ be a finite shared support, let $q_n(s) > 0$ for every n, s, and suppose Eq. 7's softmax inputs are represented by $m_n(s) = \log q_n(s) + kappa_n$ with $kappa_n$ constant in s and fixed declared weights w_n. Then* softmax$(\sum_n w_n m_n)$ *equals the log-linear pool of (6). This identifies the categorical message-combination term under those assumptions only. Independently, the project's robust server aggregator at $c = 0$ equals that log-linear pool by (7): every reweighting multiplier is* $\exp(0) = 1$*, the iteration is skipped, and the same pool code path is returned. Neither statement reproduces the complete source protocol or certifies behavior at $c > 0$.*

Corollary 6 (Closed-form Bayes recovery)**.** *With the KL divergence and the NLL loss, the generalized posterior of (1) equals the closed-form prior-times-likelihood Bayes posterior,*

$$q^*(s) \;\propto\; pi_0(s) \prod_i p(o_i|s),$$

so `generalized_posterior(KLD,NLL)` reproduces standard Bayes. Pooling those local posteriors in this project gives the log-linear pool of (6); under the assumptions of Theorem 5, that is the categorical message-combination specialization of Eq. 7, not a recovery of its complete source protocol.

The largest observed discrepancy between `robust_aggregate(robustness=0)` and `log_linear_pool` is 0 — bit-identical, since the zero-robustness branch runs the same code path — and between `generalized_posterior(KLD, NLL)` and the analytic Bayes posterior is 5.55e-17, exact to machine precision (about one ULP); both are reported in sec. 7.1, so eq. 7 and eq. 6 are verified identities rather than approximations.

The honesty contract binds at exactly this point. The recovery theorem and its corollary cover only the recovery identity and the per-agent rigorous axis of sec. 5.7; no statement *about* `robust_aggregate` transfers a bounded-influence guarantee to that divergence-reweighting, whose positive property is the `robustness` $= 0$ limit of eq. 7. A scoped no-go rejects a declared separable objective class without supplying a broader objective certificate. The per-agent influence weights that the heuristic produces under contamination are *illustrated*, not guaranteed, in fig. 13 and fig. 16; the genuine per-client FedGVI property is the rcce/AR client loss of sec. 5.7. The next subsection closes this exact gap on the server side with a *different*, objective-backed aggregator.

5.8.2 Variational aggregation with objective-backed weight control

The related server construction becomes a genuinely variational rule by replacing the heuristic's reverse-KL weight update with a forward cross-entropy update. For $c > 0$ and $\lambda > 0$, treat the consensus q and a vector of effective weights $a = (a_n)$ as joint variational parameters and define the **aggregation free energy**

$$F_\lambda(q, a) \;=\; \sum_n a_n \,\mathrm{CE}(q, q_n) \;-\; \lambda H(q) \;+\; \tfrac{1}{c}\,\mathrm{KL}_{\mathrm{gen}}(a \,\|\, w), \tag{8}$$

where $\mathrm{CE}(q, q_n) = -\sum_i q_i \log q_{n,i}$ is the cross-entropy of the consensus relative to agent n, $H(q)$ is the consensus entropy, c is the robustness, and $\lambda > 0$ is the `entropy_weight` coefficient (default $\lambda = 1.0$); $\mathrm{KL}_{\mathrm{gen}}(a \,\|\, w) = \sum_n [a_n \log(a_n/w_n) - a_n + w_n]$ is the generalized KL between the effective and base weights. Each block of F_λ has a closed-form minimizer, so alternating

$$q \;\leftarrow\; \mathrm{softmax}\Big(\tfrac{1}{\lambda}\textstyle\sum_n a_n \log q_n\Big), \qquad a_n \;\leftarrow\; w_n \,\exp\big(-c\,\mathrm{CE}(q, q_n)\big) \tag{9}$$

is exact block-coordinate descent on eq. 8 (`variational_aggregate`) for $c > 0$ and $\lambda > 0$. The implementation defines the $\lambda \downarrow 0$ endpoint separately as a deterministic tied-argmax rule; $\lambda = 0$ is not substituted into the objective or its q-update. This substitution changes both orientation and scale: $\mathrm{CE}(q, q_n) = \mathrm{KL}(q\|q_n) + H(q)$, and its common $H(q)$ term scales all raw weights, which changes the entropy of the subsequent unnormalized weighted log pool. The paired q- and a-updates in eq. 9 are exact block minimizers of the stated objective; that fact does not derive the reverse-KL heuristic. Because F is biconvex, we run the descent **multi-start** (pool, uniform, and arithmetic-mean seeds, lowest observed F among converged starts; otherwise the lowest unfinished trace is returned with `converged=False`) so a near-one-hot adversary is not left at the product-of-experts seed in the tested contamination regimes — the detail that supports the effective-weight diagnostic (sec. 11). The full derivation, the formal statement (block descent, $c \to 0$ recovery, and the raw effective-weight bound), and the numerical witnesses are in sec. 11; the empirical descent and influence collapse are shown in fig. 14 and fig. 15 and reported in sec. 7.5.2. The authoritative notation supplement records the complete objective contract in eq. 29.

This upgrades the server side from an untracked heuristic to a derived generalized-Bayes aggregation with an explicit redescending raw-weight property: a single confidently-wrong agent earns raw weight $a_n = w_n \exp(-c\,\mathrm{CE}(q, q_n)) \le w_n$ that vanishes as it diverges, whereas the naive pool grants every agent the fixed weight w_n however wrong it is. The trade is conservatism — the $-H(q)$ term makes the stationary point a maximum-entropy-biased consensus consistent with the weighted cross-entropies, so `variational_aggregate` is deliberately flatter than the product-of-experts and does *not* maximize peak point-accuracy. The two server-side rules therefore play complementary, never-conflated roles, both reported: the sharp `robust_aggregate` heuristic for accuracy under contamination (sec. 7.5.1) and the conservative `variational_aggregate` for a server-side objective with stated weight control (sec. 7.5.2). A temperature parameter $\lambda > 0$ (controlled by `entropy_weight`, default $\lambda = 1.0$) generalizes the objective to F_λ; lower λ sharpens the variational q-block toward a maximizing state for its current weighted log pool. The tempered family (sec. 11.5; objective eq. 25) recovers the full-entropy variational aggregator at $\lambda = 1.0$ and has a separately implemented deterministic tied-argmax endpoint as $\lambda \downarrow 0$; neither endpoint is guaranteed accurate and neither is an algebraic recovery of `robust_aggregate`. The effective-weight a-update is unchanged for every $\lambda > 0$, so the raw-weight bound holds over the objective-defined family.

5.9 Belief sharing: the standard aggregation corner

`belief_sharing.share_round` lifts the aggregation rule of sec. 5.8 to a colony of categorical sentinel agents. Each agent has a private sensory outcome and a local posterior over the same shared latent location; it broadcasts the posterior, not its raw observation. Following the sensory attenuation that the active-inference formulation of belief sharing imposes [Friston et al., 2024] — "agents do not hear themselves" — an agent's heard consensus excludes its own message:

$$q_{-n} \;=\; \mathrm{normalize}(q/t_n)\,, \tag{10}$$

so the round in eq. 10 implements the declared categorical colony-hive-mind mechanism. With the naive fusion rule of eq. 6, eq. 10 is the project's standard log-linear-pool consensus. Under the explicit shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, it specializes Eq. 7's message-combination term; it does not reconstruct the complete source protocol. With the server-side robust rule it yields a hive-mind that can down-weight a contaminated sentinel — an effect the contamination sweep of sec. 7.5 measures rather than assumes, and one that carries no guarantee beyond the recovery limit. The per-round diagnostics — the post-sharing belief matrix, the global consensus, and the mean surprise and accuracy against a known ground-truth state — are returned by `share_round` and drive Studies 1 and 4.

fig. 3 makes this concrete: three sentinel cards begin with different private categorical views of the nine-cell world, convert those views into local posteriors, and send only those posteriors to a fusion route. The return path is a cavity message, so the consensus heard by agent n excludes the local posterior q_n. The figure remains a protocol map rather than a new result; the empirical belief matrix and free-energy comparison remain the evidence surfaces in fig. 9 and fig. 8.

Because eq. 10 calls the aggregation rule of eq. 6 or its robust generalization, the recovery identity of eq. 7 propagates upward: a colony running `share_round` at zero server robustness is bit-identical to a colony running the project's standard log-linear-pool round. Under the qualified categorical bridge, the pool realizes only the source message-combination specialization; the robust round is a project extension, not a reconstruction of the active-inference ensemble literature [Friston et al., 2024, Heins et al., 2024]. sec. 7.2 reports that communicating colonies reach a mean variational free energy of 13.2190 nats against 16.5298 nats for incommunicado colonies across 480 seeds, with the per-agent belief matrix before and after a round shown in fig. 9 and the colony comparison in fig. 8.

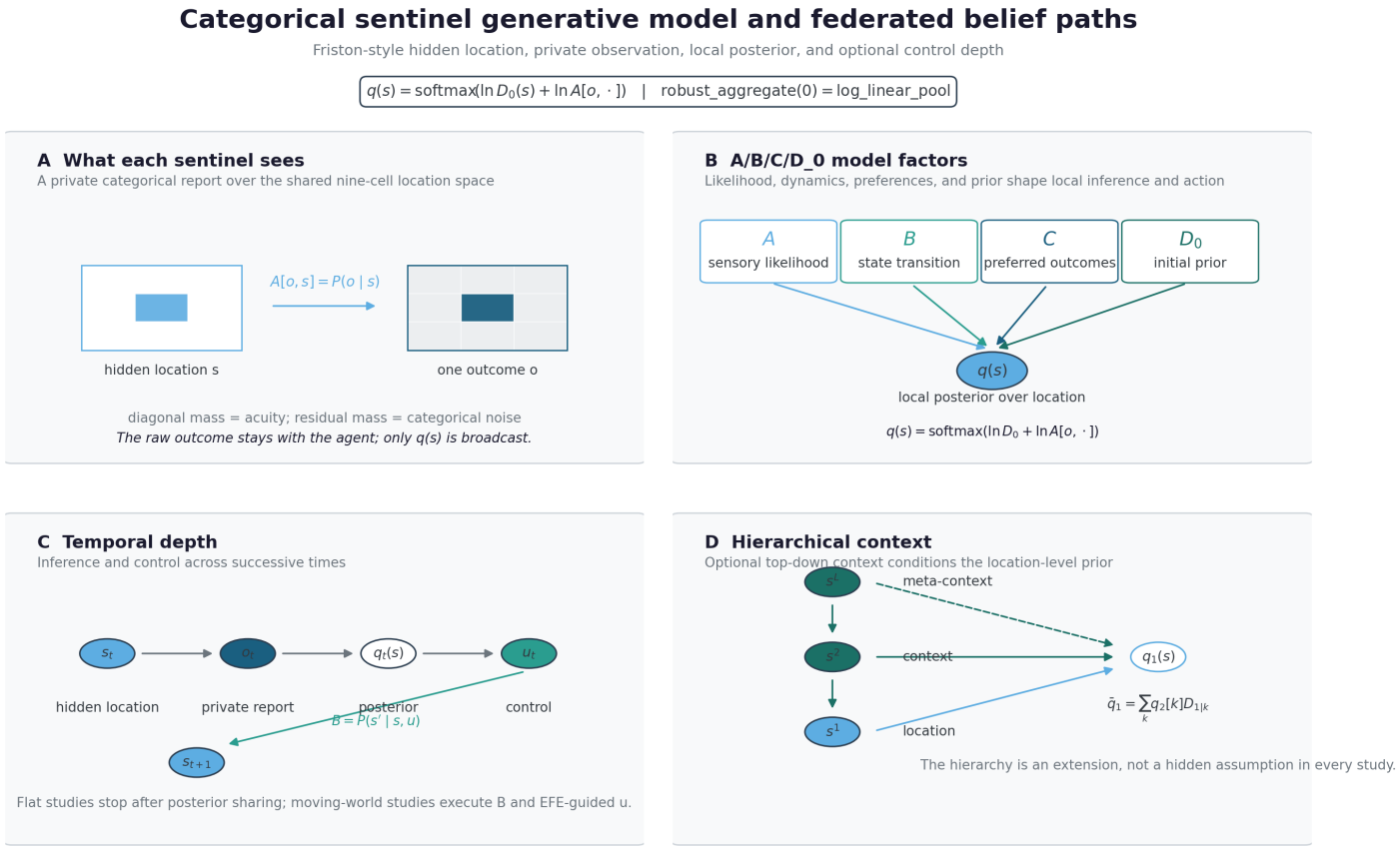
The honesty boundary of sec. 3.3 carries through the lift unchanged. The robustness that eq. 10 inherits when the colony fuses with the server-side heuristic is the divergence-reweighting device of sec. 5.8, whose positive property is the naive-recovery limit and whose declared separable objective class has a scoped no-go result; the per-agent FedGVI bounded-influence result enters the colony only through the rcce/AR client losses of sec. 5.7, under the source theorem's matching assumptions, applied inside each agent's local generalized-Bayes update. The robustness sweep in sec. 7.5 and the variational supplement in sec. 11 keep the three axes distinct. The federation transport (sec. 12.4) realizes this sharing over queue-backed worker channels; by Proposition 12, the federation bit-identity result, the consensus is bit-identical to the in-process call, so the channel adds no precision loss while leaving multi-machine network transport as future work.

5.10 Generative model: categorical states, observations, actions, and hierarchy

The 9 studies — including the contaminated-sentinel robustness sweep (Study 4) — run on one shared sentinel world (Studies 5–7 use its moving and hierarchical variants): a discrete sentinel partially-observable Markov decision process (POMDP) using the categorical world structure illustrated by Friston et al. [2024], Figures 1 and 4. We adopt the discrete-state active-inference formulation that the community has standardized around — the categorical $A/B/C/D_0$ generative model of da Costa et al. [2020], the same object the `pymdp` [Heins et al., 2022] and RxInfer [Bagaev et al., 2023] toolboxes operate on — and reimplement it in pure NumPy in `pomdp.py` so the colony, its sensors, and its dynamics are exactly the ones the analysis executes. A colony of sentinels watches a single hidden creature whose location is the shared latent factor they federate beliefs about.

The structural map in fig. 4 follows the same generative-model vocabulary while making the implementation boundary visible: Panel A shows what one agent actually sees — one noisy categorical report over the 9 possible locations — while Panel B identifies the $A/B/C/D_0$ factors that turn that

report into a local posterior. Temporal depth then describes state, observation, posterior, and control order; hierarchical depth describes conditioned priors. It is a formal schematic, not an assertion that every displayed dependency is simultaneously estimated in every study.



Schematic formalization: the agent's private sensory outcome becomes a posterior message; the server fuses messages over the shared categorical state.

Figure 4: Formal categorical generative-model schema. Source relation: source-inspired original schematic related to Friston et al. (2024), Figs. 1 and 4; estimand: categorical dependency structure; uncertainty: none. The x-axis is the dependency or role order within each panel; the y-axis positions hidden states, observations, model factors, and optional context levels. Panel A shows a hidden 9-cell location and the corresponding categorical likelihood row $A[o, s]$, making the private sensory report explicit. Panel B shows A, B, C, and D_0 feeding the local posterior $q(s)$; Panel C shows temporal state, observation, posterior, and action order; Panel D shows optional top-down conditioned priors. The equation ribbon records the implemented state-inference form and the zero-robustness recovery identity. This deterministic formal schematic contains no fitted values, empirical sample, error band, or confidence interval.

5.11 State space: one shared latent factor

The world holds one hidden factor: the creature's location on a square grid of side L, giving $n_s = L^2$ location states. Our sentinel world uses the 3×3 cardinality illustrated in Friston et al. [2024], Fig. 1, so $n_s = 9$ — the cardinality `pomdp.N_LOCATIONS` exposes and `experiment_config` carries as `n_locations`. This single location factor is precisely the latent the colony gossips about: it is the shared argument of the log-linear pool (eq. 6) and of every belief-sharing round (eq. 10). Fixing one hidden factor keeps the recovery limits of sec. 6 closed-form and exactly testable, rather than approximated.

5.12 Four categorical tensors: likelihood, transitions, preferences, priors

The generative model is the tuple (A, B, C, D_0) in the discrete active-inference convention: a categorical probability mass function is a non-negative vector summing to one, and a likelihood matrix is shape (n_o, n_s) whose columns (indexed by hidden state) are categorical.

Observation likelihood $A = P(o \mid s)$. Each sentinel observes the creature's cell through a noisy sensor. With probability `acuity` the sensor reports the true cell; the residual mass $1 - \text{acuity}$ spreads uniformly over the other $n_s - 1$ cells. With outcome cardinality $n_o = n_s$, the single location modality is one (n_s, n_s) matrix:

$$A_{os} = \begin{cases} \text{acuity}, & o = s, \\ \dfrac{1 - \text{acuity}}{n_s - 1}, & o \neq s, \end{cases} \qquad \sum_o A_{os} = 1. \tag{11}$$

The acuity in eq. 11 tunes how peaked the sensor is: high acuity gives a near-diagonal A that pins the creature; the belief-sharing study deliberately

runs the colony at the low acuity acuity = 0.55, where no single sentinel can resolve the location alone and the colony must pool evidence to do so. When a seeded generator is supplied, each sentinel's acuity is jittered by a small non-negative perturbation, so a colony carries slightly heterogeneous likelihoods while every column remains a proper pmf.

Transition tensor $B = P(s' \mid s, u)$**.** The creature moves on the grid under three control paths — `still`, `left`, `right` — so B has shape (n_s, n_s, n_u) with $n_u = 3$. `still` is the deterministic self-loop; `left` and `right` decrement and increment the grid column, saturating at the walls (a wall-adjacent move in the wall's direction is a self-loop). All three controls act on the column index alone, so the creature's row is preserved and its motion is confined to the horizontal axis of the grid — a deliberate one-dimensional control over the two-dimensional location factor. Every slice $B_{\cdot\cdot u}$ is column-normalized by construction, so the transition is a valid categorical for each action.

Log-preference C**.** The sentinel prefers to *see* the creature near the den (the center cell), encoded as a log-preference vector of shape $(n_o, 1)$ with a positive bump on the center outcome and zero elsewhere. The preferred-outcome distribution that the expected-free-energy decomposition of sec. 5.15 uses is $p_C(o) = \mathrm{softmax}(C)[o]$.

Initial prior D_0**.** The creature is believed to start at the grid center, so D_0 of shape $(n_s, 1)$ places unit mass on the center cell. D_0 enters state inference as the log-prior of the one-step variational update (eq. 12 below).

The columns-are-pmfs invariant is not assumed — it is pinned by ISC-15, which checks that every column of A and of each $B_{\cdot\cdot u}$ sums to one.

5.13 One-step variational state inference in the grid world

Given an observation o, a sentinel forms a posterior over the creature's location by a single softmax step (Friston et al. [2024], Eq. 4): the log-prior plus the additive log-likelihood message, summed over any conditionally independent modalities m,

$$q(s) \;=\; \mathrm{softmax}\Big(\ln D_0(s) \;+\; \textstyle\sum_m \ln A_m[o_m, s]\Big). \tag{12}$$

The message $\ln A_m[o_m, \cdot]$ is the row of A_m that the observed outcome selects; summing messages over modalities makes each modality an additive evidence term — the categorical product-of-experts. The companion variational free energy, the scalar eq. 12 minimizes, is

$$F[q] \;=\; \mathbb{E}_q\big[\ln q(s) - \ln D_0(s) - \textstyle\sum_m \ln A_m[o_m, s]\big] \;=\; \mathrm{KL}(q \,\|\, D_0) \;-\; \mathbb{E}_q\big[\textstyle\sum_m \ln A_m[o_m, s]\big], \tag{13}$$

reported in nats. The one-step posterior of eq. 12 is its unique minimizer, where F equals the negative log model evidence. Both live in `belief_updating.infer_states` and `belief_updating.vfe`, and the free energy of eq. 13 is the quantity the communicating-versus-incommunicado colony comparison of sec. 5.22 scores.

This inference step is not a separate mechanism bolted onto the colony: it is the L = NLL, learning-rate-1 special case of the generalized-Bayes posterior eq. 1, reusing the same locked softmax. That client identity recovers the stated categorical Bayes substrate at its trusting limits. The separate server identity in sec. 5.8 then yields the project log-linear pool under its qualified Eq. 7 message-combination bridge; together these do not recover the complete Friston protocol (sec. 6).

5.14 Hidden-state to action loop: the POMDP substrate

The categorical POMDP loop in fig. 5 separates the common latent-state substrate from the federation transport. In the sentinel interpretation, an agent is a location-sensitive observer: the hidden state is one of 9 cells, the private outcome is a noisy categorical report of that location, and the agent sends its posterior over the location rather than the report itself. The flat belief-sharing studies use the observation, posterior, and communication subset; the moving-world extension also executes transition and EFE-guided action paths. The diagram therefore gives readers the active-inference context without turning a conceptual loop into a claim that every study estimates every latent or policy quantity.

5.15 Learning stack: EFE, Dirichlet updates, and BMR

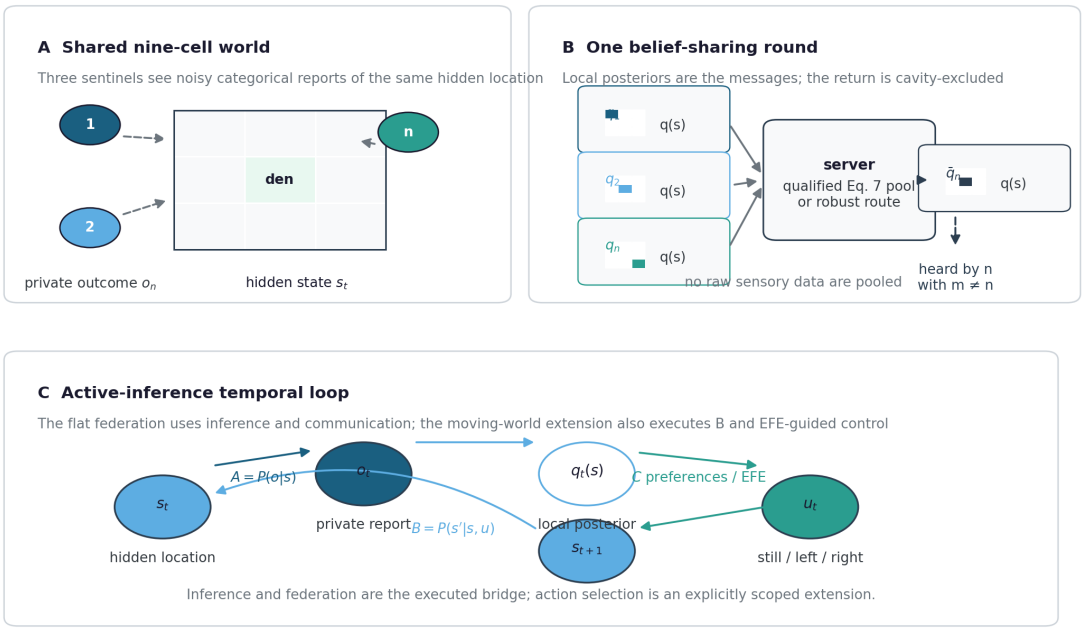
Active-inference agents do more than infer states under a fixed model: they learn the parameters of the model, score policies by expected free energy, and revise model structure. The active-inference community has standardized all three operations [Da Costa et al., 2020, Smith et al., 2022], and Friston et al. [2024] place them at the heart of the federated belief-sharing scenario. We reimplement each in closed form so the language-acquisition, expected-free-energy, and emergence studies of sec. 5.22 rest on machine-checkable quantities rather than fitted curves.

5.16 Conjugate Dirichlet learning from co-occurrence counts

A sentinel learns its observation model A by placing a Dirichlet prior with concentration a on each column and updating it conjugately from observation-state co-occurrence counts (Friston et al. [2024], their equations 9–12). One learning step adds the expected sufficient statistics for that step and reads off the column-normalized posterior mean,

$$a \;\leftarrow\; a + \mathrm{counts}, \qquad \mathbb{E}[A]_{os} \;=\; \frac{a_{os}}{\sum_{o'} a_{o's}}, \tag{14}$$

implemented in `dirichlet_learning.learn_likelihood`. Intuitively, each concentration vector is a running tally of how often each outcome was seen while the creature occupied a given state: the prior seeds that tally with pseudo-counts, every step adds the co-occurrences it witnessed, and the posterior mean is simply the tally renormalized into a categorical. Likelihood learning is therefore bookkeeping — accumulate counts, then normalize — with no iterative optimization loop. We drive eq. 14 with the expected sufficient statistics under the true model — a fixed count batch count_scale $\cdot\, A^\star$ per step, optionally jittered by a seeded generator. As the concentrations accumulate, the expected likelihood $\mathbb{E}[A]$ converges to the data-generating $A^\star$; convergence is measured by the per-column KL divergence summed over hidden states,



Model schematic, not an empirical result: panels A and B explain the Friston-style agents and messages; panel C separates the optional active-control pathway from the belief-sharing transport.

Figure 5: Sentinel-world and active-inference loop. Source relation: source-inspired original schematic related to Friston et al. (2024), Figs. 1 and 4; estimand: POMDP message-and-action sequence; uncertainty: none. The x-axis is the POMDP cycle in Panel C from hidden state through observation, posterior, action, and next state; the y-axis separates the shared-world, belief-sharing, and temporal-loop panels. Panel A shows three agents viewing the same 9-cell hidden world through private, noisy categorical observations; raw observations remain local. Panel B shows those local posteriors entering a log-linear-pool or robust server and returning as a cavity-excluded consensus. Panel C gives the POMDP cycle from hidden state s_t through observation o_t, posterior $q_t(s)$, action u_t, and transition to s_{t+1}, with A, B, and C marking the likelihood, transition, and preference factors. The flat studies execute the inference-sharing branch; the moving-world extension also executes transitions and EFE-guided actions. This is a deterministic model schematic, not an uncertainty-bearing empirical result.

$$\mathrm{KL}(A^\star \,\|\, \mathbb{E}[A]) \;=\; \sum_s \sum_o A^\star_{os} \,\ln \frac{A^\star_{os}}{\mathbb{E}[A]_{os}}, \tag{15}$$

which decreases monotonically toward zero — the standard-Bayes / KL fixed point. The learned likelihood always has full support (the Dirichlet prior is strictly positive), so eq. 15 is finite throughout. ISC-17 pins the monotone-decreasing KL trajectory, and the language-acquisition study of sec. 5.22 reports the descent of eq. 15 across 24 steps.

The implementation also carries the η forgetting hyperprior of Friston et al. [2024], their equation 12: before each conjugate addition the running concentration is decayed so the *total* concentration mass saturates at η rather than growing without bound, modeling an agent that stays adaptable instead of becoming infinitely confident. With η unset the classical unbounded accumulation of eq. 14 is recovered.

5.17 Expected free energy as the action-selection objective

A sentinel scores a candidate policy π by its expected free energy $G(\pi)$, which the active-inference formulation decomposes two equivalent ways (Friston et al. [2024], their equation 2): a cost view of risk plus ambiguity, and a value view of pragmatic plus epistemic value. The two views are the same scalar rearranged, stated as eq. 18 and pinned to a zero residual by the algebraic identity eq. 19 in sec. 6. We compute every term in closed form from the categorical model (A, B, C, D_0) in `expected_free_energy.decompose`:

- *Risk* is $\mathrm{KL}(q(o \mid \pi) \,\|\, p_C(o))$, the deviation of the policy-predicted outcomes from the preferred-outcome pmf $p_C(o) = \mathrm{softmax}(C)[o]$; write $q_\pi(o) := q(o \mid \pi)$ for this scored-policy outcome predictive, used in the remaining terms.
- *Ambiguity* is the expected likelihood entropy $\mathbb{E}_{q(s)}[H[p(o \mid s)]]$, the outcome uncertainty given the state.
- *Pragmatic value* is the expected log-preference $\mathbb{E}_{q_\pi(o)}[\ln p_C(o)]$ — the utility, exploitation term.
- *Epistemic value* is the state-outcome mutual information $H[q_\pi(o)] - \mathbb{E}_{q(s)}[H[p(o \mid s)]]$ — the expected information gain that drives exploration.

Because there is no sampling, the identity of eq. 19 holds to floating-point tolerance; ISC-19 (`expected_free_energy`) pins the residual of the decomposition to zero and pins each term's semantics independently (deterministic likelihoods give zero ambiguity; uninformative likelihoods give zero epistemic value; preference-matched predictions lower risk). fig. 7 visualizes the additive cost view and the signed pragmatic/epistemic waterfall terminating at $G(\pi)$ — a deterministic identity (Proposition 7), not a fitted result.

5.18 Bayesian model reduction for structure emergence

Sentinels also revise model *structure*. Bayesian model reduction (BMR) scores whether a reduced model — for example one that prunes a redundant location column by shrinking its concentration toward zero — has more evidence than the full model, *without re-running inference* (Friston & Penny via the post-hoc model optimization lineage [Friston and Penny, 2011]; the same Beta-function identity is their equation 13 (Friston et al. [2024])). Because the likelihood is shared, the reduced posterior is available in closed form, reduced_post = post + reduced_prior − prior, and the change in (negative) variational free energy is a difference of log multivariate Beta functions,

$$\begin{aligned} \Delta F &= \ln B(\text{prior}) + \ln B(\text{reduced\_post}) - \ln B(\text{post}) - \ln B(\text{reduced\_prior}), \\ \ln B(a) &= \textstyle\sum_k \ln \Gamma(a_k) - \ln \Gamma(\sum_k a_k), \end{aligned} \tag{16}$$

computed in `bayesian_model_reduction.reduce`. A positive ΔF in eq. 16 means the reduced model carries more evidence — the pruned structure was redundant and should be adopted; a negative ΔF means the reduction destroyed support the data require. When the reduced prior equals the prior the score is identically zero in exact algebra, a zero point the suite pins to machine precision (ISC-20).

The emergence study of sec. 5.22 uses this operation over $n = 4$ candidate states. It contrasts a redundant reduction ($\Delta F = 3.68$ nats; adopt) with a supported one ($\Delta F = -27.67$ nats; reject) in fig. 11. This fixed-candidate algebraic comparison is deterministic, so it has no resampled sample or bootstrap interval.

5.19 Contamination models: declared failure modes for belief fusion

Robust belief fusion only earns its keep when some agents are wrong. The active-inference community has built ensembles that coordinate by sharing beliefs and observations [Friston et al., 2024, Heins et al., 2024, Albarracin et al., 2022, Kaufmann et al., 2021], but it has assumed those beliefs are trustworthy in the cited modeled protocols: fusion is treated as exact-Bayes pooling of well-calibrated reports. The robust-Bayes and federated-learning literatures [McMahan et al., 2017, Ashman et al., 2022, Mildner et al., 2025b] have, in turn, studied robustness to corrupted clients under their declared settings, but outside the generative-model-bearing POMDP setting. This section defines the corruption process that lets us test fusion robustness inside the active-inference colony — the experimental complement of the robust aggregation rule of sec. 5.8.

5.20 Corruption process for adversarial belief broadcasts

In the sentinel world a healthy sentinel reports a well-calibrated categorical over the creature's location; a contaminated one reports something corrupted. `contamination.contaminate` manufactures the corrupted reports. Every corruption is a convex mixture of the agent's belief b with a corruption target t, governed by a single rate $r \in [0, 1]$,

$$\tilde{b} \;=\; (1 - r)\, b \;+\; r\, t, \tag{17}$$

so the experiments sweep exactly one knob. The convex form of eq. 17 gives a clean limit and is the anchor of the suite (ISC-26): at $r = 0$ every corruption kind returns the input belief unchanged, so contamination is a strict, continuous departure from the uncorrupted Friston belief-share —

never a discontinuity. This section defines the three core corruption targets t, each capturing a distinct failure of a federated agent. Geometrically the three are three landmarks of the probability simplex — a wrong vertex (`confident_wrong`), the flat centroid (`uniform`), and a random interior point (`label_noise`) — so the mixture of eq. 17 drags an honest belief toward a qualitatively different destination in each case. Two further mechanisms (`byzantine` and `drift`) extend the same convex-mix contract and are introduced in the extended-methods supplement (sec. 12.2).

`confident_wrong` — the adversarial sentinel. This is the lookout that points to one wrong cell and insists on it with total certainty. The target is a one-hot spike on a wrong state, $t = \text{onehot}(s_{\text{wrong}})$, so $\tilde{b}$ is mixed toward a confident, mistaken delta. Callers choose s_{wrong} explicitly; the verdict sweep of sec. 5.22 fixes it once per colony as the state diametrically opposite the true state on the location grid, held constant across the entire rate sweep, rather than deriving it from the agent's current belief. At $r = 1$ this is a pure delta on the wrong cell. This is the saboteur that is *sure* and *mistaken*: exactly the agent that robust aggregation must reject.

`label_noise` — the miscalibrated sentinel. This is the lookout with a scrambled sensor: it is not lying toward any particular cell, only diluting every honest report with the same fixed sprinkle of noise. The target is a fixed noisy categorical drawn once from a Dirichlet(1) (a random but valid pmf), modeling a sentinel whose report is partly random rather than adversarial. Because the noisy target is drawn once and then held fixed across the rate sweep, the corruption has no direction to exploit and no single cell to veto — the robust pool meets diffuse degradation, not a targeted attack.

`uniform` — the apathetic sentinel. This is the lookout that shrugs: it has lost track of the creature and calls every cell equally likely. The target is the maximum-entropy uniform pmf $t = (1/n_s)\mathbf{1}$, modeling a saturated sentinel that has lost all information. At $r = 1$ it reports uniform, contributing no evidence to the pool rather than actively pulling it toward a wrong cell.

All three share the contract of eq. 17 and require an explicit seeded generator — `label_noise` uses it to draw the noisy target — so every contaminated report is reproducible. The grid of rates the sweep uses, $\{0, 0.225, 0.45, 0.675, 0.9\}$, deliberately stops below the pure-veto limit $r = 1$, where a fully-confident wrong delta forces *every* pooling rule's accuracy to zero and the robust-versus-naive contrast vanishes.

5.21 How contamination meets the three robustness axes

A contaminated report feeds the colony in distinct places, and the honesty contract of sec. 3.3 turns on keeping them separate.

At the *server* (the aggregation step of sec. 5.8) a contaminated belief enters `robust_aggregate`, the iteratively-reweighted pool that discounts each agent by $\exp(-c\,\mathrm{KL}(q_n \,\|\, q))$. A confidently-wrong agent sits far from the emerging consensus, earns a small effective weight, and is suppressed. This is the *heuristic* axis: its only proven property is that at $c = 0$ it recovers the project's naive log-linear pool exactly (eq. 7, Theorem 5). Under the qualified bridge of sec. 5.8, that pool specializes Eq. 7's message-combination term rather than the complete source protocol. The robustness-sweep figures (fig. 12, fig. 13) illustrate this heuristic's behavior — including the per-agent influence weights that drop the saboteurs — but they do not certify a per-agent guarantee.

At the *client* (the per-agent generalized-Bayes update of sec. 5.7) contamination is what the bounded β-loss (eq. 4) and rcce-loss (eq. 5) are designed to survive: a single corrupted observation with $p(o) \to 0$ drives the unbounded NLL to dominate the posterior, whereas the bounded losses cap its influence. This is the source-theorem-backed axis: the FedGVI guarantee [Mildner et al., 2025b] is inherited only under the source theorem's matching loss, divergence, and regularity assumptions. The federated logistic-regression baseline of sec. 5.22 applies this same client mechanism to flipped-label contamination (fig. 16); it is the conjugate Bernoulli analogue of the categorical client update, and its robustness is the per-client loss, not the server reweighting.

No figure, statistic, or sentence in this manuscript grants the server-side heuristic the per-client bounded-influence guarantee; contamination is the common stressor against which the three axes are kept distinct.

5.22 Experimental design: studies, estimands, determinism, and power

The generative model of sec. 5.10, the learning operators of sec. 5.15, and the corruption process of sec. 5.19 are exercised by 9 studies, including the contaminated-sentinel robustness sweep (Study 4). The shared configuration (seed budget, colony size, contamination rates, divergences, trial counts, and the statistics settings) is read from `experiment:` in `manuscript/config.yaml`; the remaining per-study parameters are tested code defaults in `src/fedference/experiments/`. No value is hard-coded in the manuscript, and each token below resolves to the same configuration the code executed.

5.23 Determinism through fixed seeds and generated variables

All stochastic steps draw from explicitly seeded generators (`np.random.default_rng`); the global `np.random` state is never touched. The single-run studies use the first configured seed (0), and the across-seed studies enumerate the deterministic seed list $0, \ldots, n_{\text{seeds}} - 1$. Re-running with the same seed reproduces every number in the results bit-for-bit, so the bootstrap confidence intervals and paired-test p-values of sec. 5.27 are themselves deterministic functions of the seed.

The figure layer follows the same provenance rule. Captions are written to be self-contained, with axes, resampling units, deterministic runs, truncated axes, and error-band status disclosed in the caption rather than left to inference [Rougier et al., 2014, Midway, 2020]. This is why several results figures state "single deterministic run" or "no error band" even when the inferential evidence appears in an adjacent table.

5.24 Study suite and contamination sweep

Studies 1–3 implement reduced categorical protocols that are source-mechanism analogues of the belief-sharing, language-acquisition, and model-reduction mechanisms discussed by Friston et al. [2024] on the sentinel POMDP; they are not exact source-protocol figure replications. The sweep adds the FedGVI robustness contribution [Mildner et al., 2025b]. Unless a study specifies otherwise, the global defaults are $n_{\text{seeds}} = 480$ independent seeds and $n_{\text{trials}} = 960$ matched trials per condition.

Table 1: Per-study configuration, read from `experiment:` in `config.yaml` and surfaced as manuscript tokens by `src/manuscript_variables.generate_variables`. The sample sizes carried into the statistics — the across-seed belief-sharing sample ($n = 480$ seeds), the language trajectory (25 ordered points summarized over $n = 480$ independent seeds), and the paired robustness trials ($n = 960$ per condition) — are reported with their respective results.

Study	What it measures	Key parameters
1 — Belief sharing	Communicating vs. incommunicado colony free energy	`n_agents = 7`, `acuity = 0.55`
2 — Language acquisition	Dirichlet-learning KL descent (eq. 15)	`num_steps = 24`
3 — Emergence / BMR	Reduced-vs-full model evidence (eq. 16)	candidate states `n = 4`
4 — Robustness sweep	Robust vs. naive consensus under contamination	`n_agents = 7`, `n_contaminated = 2`
5 — Disjoint-FOV moving world	Communication necessity with non-overlapping fields of view	`n_agents = 3`, `fov_width = 2` (Supplement, sec. 13.7)

Study 1 — belief sharing. A colony of 7 sentinels at the deliberately low acuity 0.55 each infer the creature location (eq. 12) and share beliefs through the log-linear pool (eq. 6). We compare a *communicating* colony against an *incommunicado* one of the same size and seed, scoring each by the mean variational free energy of eq. 13. Two protocol details: state inference in this study substitutes a flat (uniform) prior for D, and each belief's free energy is scored against the pooled evidence of all agents' observations (the disclosure carried in sec. 7.2). The across-seed sample is $n = 480$ seeds, one colony-mean free energy per seed (fig. 8, fig. 9).

Study 2 — language acquisition. Each configured seed runs one sentinel trajectory using the conjugate Dirichlet update (eq. 14) over 24 steps. We record the KL descent of eq. 15 at each ordered step, giving 25 points per trajectory and $n = 480$ independent seed trajectories for the pointwise interval in fig. 10.

Study 3 — emergence. A redundant model reduction and a supported one are scored by the BMR free energy of eq. 16 over $n = 4$ candidate states (fig. 11).

Study 4 — robustness sweep. The sweep varies two factors on a colony of 7 sentinels of which 2 are saboteurs (`confident_wrong`, sec. 5.20):

- **Contamination rate** over $\{0, 0.225, 0.45, 0.675, 0.9\}$ — the convex-mix weight of eq. 17 toward the confident-wrong delta.
- **Server robustness setting**, named with FedGVI client-loss/ divergence vocabulary for cross-reference only $\{KLD, RKL, AR, beta, rcce\}$. `KLD` is the non-robust Friston / standard-Bayes baseline (server robustness 0, eq. 7) and serves as the design's negative control: the recovery identity guarantees it reproduces the naive log-linear pool exactly, so every robust-versus-naive contrast is scored against a comparator that is provably the un-robustified server rather than a separately tuned competitor. The remaining labels $\{RKL, AR, beta, rcce\}$ each select a fixed `robust_aggregate` down-weighting constant (the executed mapping is KLD (c=0.00), RKL (c=1.50), AR (c=1.30), beta (c=1.70), rcce (c=1.60); defined in `fedference.experiments._common`). None of these labels invoke the client-side `generalized_posterior` update of sec. 5.3 or the divergence family of sec. 5.6; this sweep therefore exercises only the server-side heuristic axis of sec. 3.3 (`robust_aggregate`), never the per-agent rigorous axis. The executed per-divergence down-weighting strengths are fixed constants defined in `fedference.experiments._common` and recorded in the run reports.

The headline verdict pairs 960 independent trials at the fixed contamination rate 0.800 — heavy contamination that degrades the naive pool while staying below the pure-veto cliff (fig. 12, fig. 13). Each trial contributes one matched (naive, robust) accuracy pair, so the replication unit is the trial and the estimand is the within-trial accuracy difference at that single rate. A complementary federated logistic-regression baseline applies the same client-side robust loss to flipped-label contamination, isolating the rigorous axis (fig. 16).

Study 5 — disjoint-FOV moving world. This extension (sec. 13.7) places 3 agents on a 2-slot disjoint-FOV track to test the necessity of communication when agents cannot observe the same positions. Isolated agents are compared against EFE-guided communicating agents on accuracy across the moving sentinel's trajectory. Five structural extension studies (Studies 5–9, Supplementary sections) build on the same POMDP substrate and are described there: the moving disjoint-FOV sentinel (sec. 13.7), the 2-level hierarchical POMDP (sec. 13.9), the N-level extension (sec. 13.11), the 2-D sensitivity sweep (sec. 13.13), and parameter recovery (sec. 14).

5.25 Sample size and prospective statistical power

The verdict design answers a deliberate question: pair *many* trials at *one* high contamination rate rather than spread few trials across the rate curve. A matched-pairs Wilcoxon test gains power from the number of matched pairs, so concentrating $n = 960$ paired trials at the single rate 0.800 gives the test the resolution to detect the robustness effect; scattering the same budget across 5 rates would dilute every contrast. The across-seed studies are powered separately, with $n = 480$ seeds for belief sharing and $n = 480$ independent seed trajectories for language acquisition. The 25 ordered learning points are repeated measures within each seed, not additional independent replicates, so the language interval does not count time points as samples. The structural-extension and cross-study summary tier uses $n = 128$ independent seeds and $n = 40$ matched trials per contamination rate for its robustness row. Trials and clients are nested within a seed and are reduced before across-seed inference; they are not additional independent replicates.

The bounded red-team review grid is a separate source-bound analysis profile. It uses 160 deterministic seed replicates, with 24 trials nested within each seed and scenario/rate cell, and retains the registered rates 0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 across the finite attack union clean, confident wrong, permutation, byzantine, drift, label noise, uniform. Its independent unit is the configured seed within a declared cell; cells that share design structure are not treated as independent worlds. Robust operating points, method settings, and rate profiles are fixed before the review run. The grid reports selection-free contrasts and keeps clean, uniform, label-noise, permutation, confident-wrong, Byzantine, and drift controls visible. Completion of this finite review does not close server-heuristic characterization, leakage-free calibration, external-data, or protocol-reconstruction phases.

We do not merely assert adequacy — we compute the design power implied by the observed effect. For the headline robust method (RKL) the observed-effect design power of the paired Wilcoxon at the run's $n = 960$, computed at $\alpha = 0.05$ against the directional alternative `greater` (robust accuracy exceeds naive), is 1.0000. The power approximation uses the deterministic noncentral-normal approximation of sec. 5.27, deflated by the Wilcoxon's Pitman asymptotic relative efficiency, so it is approximately calibrated for the signed-rank test actually run (exact under a normal-shift alternative; the power computation is one-sided while the reported p-values are two-sided). To bound a confirmatory replication, the prospective sample size needed to reach the target power 0.80 at the observed effect is $n = 5$ matched trials — the explicit sample-size budget a follow-up study should adopt. These power quantities characterize the *server-side* aggregation contrast; per the honesty contract of sec. 3.3 they do not certify the per-client β/rcce guarantees, which are pinned by the locked core (sec. 6) rather than by these aggregation-level statistics.

5.26 Software environment and configuration fingerprint

The FedGVI core, the POMDP studies, and the logistic-regression baseline are pure NumPy / SciPy — no GPU and no network. The deterministic-MLP neural complement (sec. 7.6) additionally uses PyTorch (CPU); the analysis pipeline executes it and emits its numbers as tokens exactly like every other result when the `torch` optional extra is installed, and otherwise records a skipped status with unavailable-value sentinels instead of silently fabricating neural results. Versions and platform — including the PyTorch version used for the MLP complement — are recorded automatically in sec. 10.

5.27 Statistical protocol: matched comparisons, intervals, and bounded claims

No "robust beats naive" claim is written before a statistical test produces it (Algorithm Gate I). Where the active-inference community has typically reported belief-fusion outcomes as single illustrative runs, we treat every headline contrast as a paired hypothesis test with effect-size estimation, multiple-testing deflation, confidence intervals, and an observed-effect design-power calculation — the reporting discipline expected by robust statistics and applied inference [Huber and Ronchetti, 2009, Efron and Tibshirani, 1993, Koehler et al., 2009, Morris et al., 2019, Loy and Korobova, 2021, Nakagawa and Cuthill, 2007, Benjamini and Hochberg, 1995, Wasserstein and Lazar, 2016]. The protocol lives in `statistics.py`, sits at the analysis tier above the locked FedGVI core, and emits every number the results sections report; nothing is hard-coded (ISC-30).

5.28 Paired comparison and standardized effect size

The headline claim is *paired*: across matched scenarios — same seed, same contamination — does the robust aggregator raise consensus accuracy over the naive log-linear pool? For the expanded review grid, the inferential unit is the seed: trials are nested within a seed/cell and are averaged before seed-level contrasts. The primary robustness sweep also retains a trial-level paired diagnostic, but its matched trial replicates are nested within one fixed seeded world and are not independent worlds. The seed-level review-grid estimand is therefore the robust-minus-naive accuracy difference after the declared trial reduction, not a contrast of two independently drawn group means. Each configured robust method is retained in every review-grid rate panel and has its own seed-level contrast, interval, and test; no pooled-selected curve or pooled-selected inferential member enters that grid. The naive comparator is the project's log-linear pool eq. 6. Under the shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, it specializes the message-combination term of Friston et al.'s Eq. 7 [2024], rather than reconstructing the complete source protocol. We test it with the matched-pairs Wilcoxon signed-rank test [Wilcoxon, 1945] (`statistics.paired_test`, ISC-28), interpreted under the usual signed-rank conditions: seed-level paired differences are independent across the declared seed schedule, while the primary sweep's trial-level diagnostic remains conditional on its fixed world; the differences are ranked after zero differences are removed, and the null is a symmetric distribution of paired differences around zero rather than an equality-of-means claim [Fay and Proschan, 2010]. This is the right comparison for bounded, often non-Gaussian accuracy deltas, but it is not assumption-free. The test reports the primary matched-pairs rank-biserial effect $r_{\mathrm{rb}} = (T^+ - T^-)/(T^+ + T^-)$. We retain the monotone secondary transform $d_{\mathrm{eq}} = 2r_{\mathrm{rb}}/\sqrt{1 - r_{\mathrm{rb}}^2}$, labeled a *rank-biserial-derived d-equivalent*, never raw Cohen's d and never a replacement for the mean contrast. When r_{rb} saturates at ± 1, the transform diverges; tables print a signed saturation marker rather than a finite million-scale value. The primary-sweep headline display method RKL has rank-biserial effect 1.0000, d-equivalent saturated (r=+1) (large), with the paired mean difference and bootstrap interval reported alongside it.

5.29 Bootstrap interval estimates

Every inferential mean — colony free energy, learning-curve KL, per-method accuracy, and the robust-minus-naive accuracy difference — carries a 95% percentile-bootstrap confidence interval [Efron and Tibshirani, 1993] from `statistics.bootstrap_ci`, resampled from the declared unit. For the review grid, the unit is the seed after the nested trials have been reduced; for the primary fixed-world sweep, the interval is explicitly trial-level and conditional; for single-seed diagnostics, it is descriptive or trial-level. The single-colony mechanistic rate table is $n = 1$ per cell and is descriptive, not an inferential mean; its companion profile uses the declared nested design. This separation follows simulation-reporting guidance to declare the estimand and the independent Monte Carlo unit [Morris et al., 2019]. The interval quantifies variation over the declared resampling unit (seed, recorded step, or matched trial) and is conditional on this simulation design, not on unmodeled real-world deployment uncertainty or alternative contamination models. The headline mean robust-minus-naive accuracy difference is 0.0846 with 95% CI $[0.0821, 0.0873]$, and the accuracy at the verdict rate is 0.9021 for the naive pool (95% CI $[0.8993, 0.9049]$) against 0.9867 for the most accurate robust member (95% CI $[0.9865, 0.9869]$).

For every multi-seed summary we also report the Monte Carlo standard error (MCSE $= s/\sqrt{n_{\mathrm{seed}}}$) and an approximate two-sided minimum detectable effect (MDE) at the configured target power. These are precision diagnostics over independent seeds, not claims about sampling a real population. The MDE uses a normal approximation conditional on the observed seed-level standard deviation [Koehler et al., 2009]; it is reported alongside the non-parametric interval rather than used to replace the paired test.

5.30 Multiple-testing deflation by BH-FDR

The sweep compares each robust divergence against the naive pool, so an uncorrected $p < 0.05$ across the family would manufacture false discoveries. We control the false-discovery rate with the Benjamini-Hochberg step-up procedure [Benjamini and Hochberg, 1995] (`statistics.bh_fdr`, ISC-29) at $\alpha = 0.05$. The verdict family owns one robust-versus-naive contrast per robust method at the predeclared verdict rate; each rate table owns its own within-method rate family. Families are not pooled across figures or across the review-grid cells. The review-grid families retain all configured method contrasts and do not derive a selected member by pooled mean. The procedure returns both the rejection mask and the monotone BH q-values. BH controls expected false discovery proportion within the stated family, not the family-wise probability of any false positive; the manuscript therefore states the family each table uses. The positive-contrast rule is strict and conjunctive: a method is a BH-rejected positive contrast *iff* BH rejects its null **and** its rank-biserial effect is positive. This is a statistical decision for the named family, not a unique scientific winner. The primary-sweep headline display method has raw $p = 1.11 \times 10^{-158}$, deflating to BH $q = 1.11 \times 10^{-158}$.

5.31 Prospective power analysis for the verdict rate

We report not only *that* the verdict rejected the null but the observed-effect planning power implied by the run and the number of pairs a confirmatory run could budget. The observed-effect design power of the headline paired Wilcoxon at the primary sweep's matched-trial unit, $n_{\text{trial}} = 960$, $\alpha = 0.05$, alternative `greater`, is 1.0000; the prospective sample size for the target power 0.80 at the observed effect is $n_{\text{trial}} = 5$. The estimator is the deterministic noncentral-normal approximation to the matched-pairs t-test, deflated by the Wilcoxon's Pitman asymptotic relative efficiency $3/\pi$, so the reported power is approximately calibrated for the signed-rank test the harness runs (the deflation is exact under a normal-shift alternative); note the power computation is directional for planning, while the reported p-values are two-sided. This is an observed-effect planning approximation conditional on the observed effect size, not independent evidence for the result and not a confirmatory power guarantee.

5.32 Reporting tables and the honesty boundary

The results sections render five statistics tables straight from these tokens: the per-rate accuracy sweep (tbl. 3), the robust-versus-naive verdict (tbl. 5), the standardized-effect verdict with observed-effect design power and prospective n (tbl. 6), the per-method accuracy with bootstrap CIs at the verdict rate (tbl. 7), and the per-contamination-rate paired tests (tbl. 4). Every cell is a generated token, never hand-typed.

The honesty contract binds at exactly this tier. The effect-size, CI, and power enrichment above decorates the *server-side* `robust_aggregate` divergence-reweighting contrasts only. It does not certify the per-client β/rcce generalized-Bayes (FedGVI) guarantees, which are pinned by the locked core (Proposition 4 of sec. 6) rather than by these aggregation-level statistics — the three robustness axes of sec. 3.3 kept distinct.

5.33 Computational complexity and scaling diagnostic

The release now reports computational complexity from the implementation itself. Let N be the number of agents, S the number of categorical states, I the solver-iteration budget, B the number of variational starts, and M the number of conditionally independent observation modalities. The dominant dense work and the storage actually retained by the current NumPy paths are:

operation	dominant time order	retained or peak storage
log-linear pooling	$\Theta(NS)$	$\Theta(NS)$
iterative robust pooling	$\Theta(INS)$	$\Theta(NS + IS)$
objective-backed variational pooling	$\Theta(BINS)$	$\Theta(NS + IS + BS)$
self-excluding naive sharing	$\Theta(N^2S)$	$\Theta(NS)$
self-excluding robust sharing	$\Theta(IN^2S)$	$\Theta(NS + IS)$
one-step state inference	$\Theta(MS)$	$\Theta(S)$
server round, excluding queue/network wait	$\Theta(NlogN + INS)$	$\Theta(NS)$

These are dominant interaction counts, not hardware-independent FLOP totals. The N^2 sharing term is material: with sensory attenuation enabled, one round computes one global pool and one leave-one-out pool per agent. The iterative server rules additionally retain their returned per-iteration histories, which is why their storage rows include IS. The server row includes worker-ID sorting and dense aggregation; incoming serialized belief volume is linear in NS, while queue and network latency are deliberately outside this local-compute accounting. The local path serializes one consensus-plus-weight result; if physical broadcast bytes are counted per recipient, that outgoing volume additionally has an N^2 term because each result carries the N agent weights.

The accompanying seeded benchmark measures the real public call paths on the declared grids $N \in \{4, 8, 16, 32, 64\}$, $S \in \{256, 512, 1024, 2048, 4096\}$, self-excluding sharing $N \in \{4, 8, 16, 32\}$, and $M \in \{1, 2, 4, 8\}$. The fixed dimensions are $N = 256$, $S = 64$ for aggregation, and $S = 16384$ for state inference; direct aggregation and server timings use $I = 6$, while the public robust self-excluding sharing path uses its solver budget $I = 32$; the default variational path uses $B = 3$. Each grid point is warmed up 1 time(s) and measured 5 time(s), with median time plotted and the observed repeat range shown as a min–max bar. The fixed input seed is 20260728 on the *arm*64 machine using Python 3.13.11 and NumPy 2.4.2.

The measured log–log slopes are descriptive checks of the expected orders, not performance guarantees: agent-axis slopes are 0.89 (log-linear), 0.94 (iterative robust), 0.71 (variational), 1.59 (naive self-excluding sharing), and 1.93 (robust self-excluding sharing); state-axis slopes are 0.42, 0.40, and 0.41; the modality-axis inference slope is 0.67. The slope fit is a timing diagnostic on this machine, not an inferential test and not evidence that the same constants hold under another BLAS, accelerator, process topology, or distributed network. A finite grid can also yield a sublinear fitted slope when validation, allocation, cache, and interpreter overheads are material; the implementation-derived order is the governing claim, not equality between a finite-grid slope and its exponent.

Figure fig. 6 visualizes the implementation-derived orders and the corresponding finite-grid timing diagnostic.

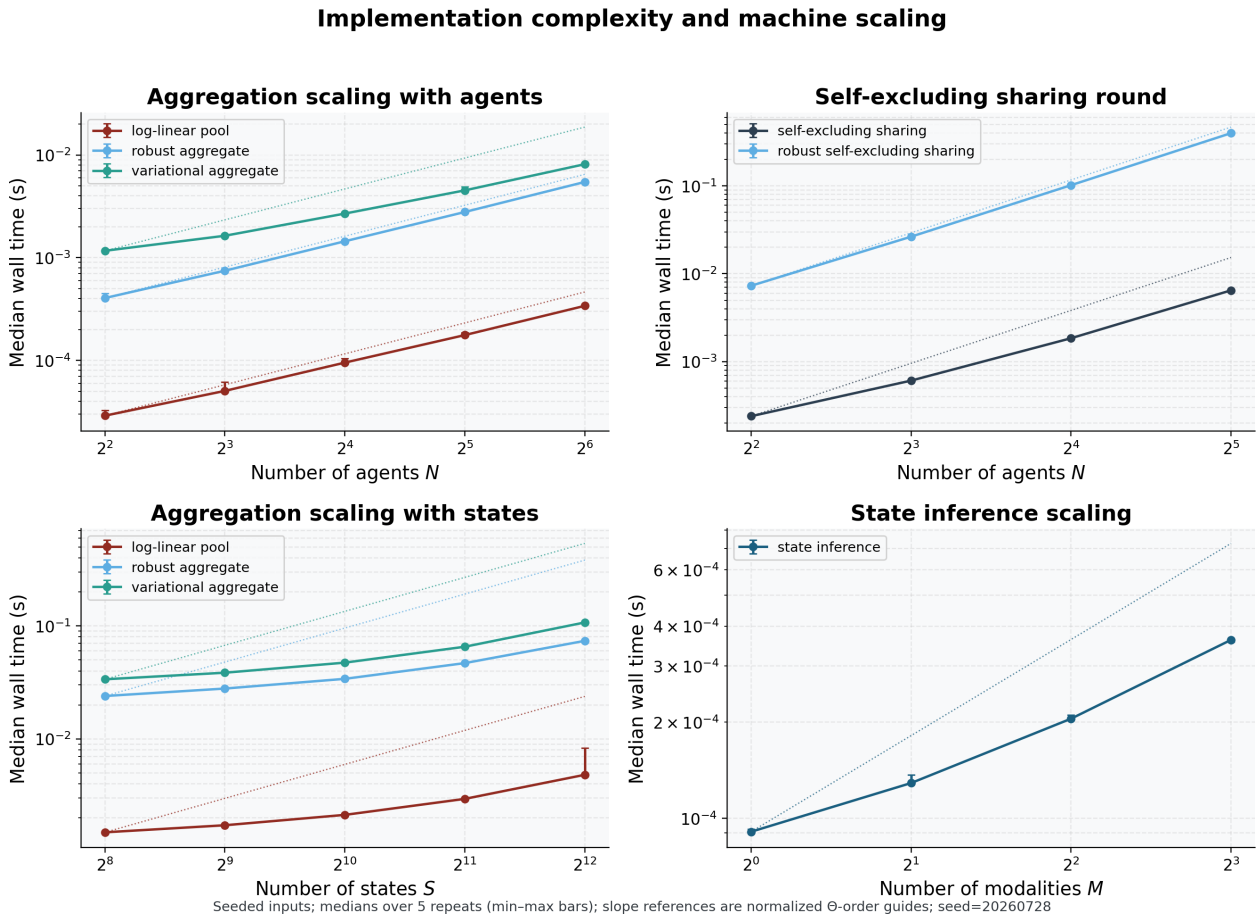


Figure 6: Implementation-derived complexity and seeded machine-scaling diagnostic. Source relation: original project computational-complexity diagnostic; estimand: median wall-clock time of the real categorical aggregation, naive and robust self-excluding sharing, and state-inference call paths as the declared dimension changes; uncertainty: min–max span over the repeated timings, not a confidence interval; replication unit: fixed seeded input at each grid point with the declared timing repeats. The x-axis is the varied agent, state, or modality dimension, and the y-axis is median wall-clock time in seconds. The panels show agent scaling for the aggregation rules, naive and iterative-robust N^2 leave-one-out sharing, state scaling for the aggregators, and modality scaling for one-step inference. Dotted lines are normalized Θ-order guides from the implementation-derived accounting; they are not fitted claims. The experiment ran on *arm*64 with Python 3.13.11, NumPy 2.4.2, seed 20260728, and 5 measured repeat(s) after 1 warmup(s).

6 Formalism: recovery limits, EFE, and tempered aggregation

The primitives of sec. 5 are governed by a compact set of machine-checkable identities. Their counter is monotone across the methods and this section: Definitions 1 and 2 (posterior and cavity / PVI update, sec. 5.3); Lemma 3 (Rényi KL limit, sec. 5.6); Proposition 4 (β-loss and rcce NLL limits, sec. 5.7); Theorem 5 and Corollary 6 (belief-sharing and Bayes recovery, sec. 5.8); and Proposition 7 below (expected-free-energy identity). Each carries a tested residual. None grants a bounded-influence guarantee to the server-side `robust_aggregate` heuristic; that boundary is stated in sec. 3.3.

6.1 Recovery limits as the proof surface

The recovery limits separate client and server claims. The divergence and loss limits recover the standard-Bayes client update; the independently tested `robust_aggregate(robustness=0) == log_linear_pool` identity recovers the project's standard server pool. Under the explicit shared-support, posterior-log-potential, and fixed-weight assumptions in sec. 5.8, that pool is a categorical specialization of Eq. 7's message-combination term, not recovery of the complete source protocol [Friston et al., 2024]. We collect the five residuals that pin those limited claims, each emitted by the test-suite and reported in sec. 7.1, never hardcoded. Read each row of the table below as a triple: the robust primitive, the trusting knob value at which it must collapse onto its standard-Bayes or project-local counterpart, and the tested residual measuring whatever gap survives at that value. The rows differ in what kind of check they are. The divergence and loss rows evaluate inside the implementation's closed-form switch band, so their zeros are exact branch identities — guaranteed by construction, not measurements that could have come out otherwise. Their genuine falsifiers are the *off-switch* convergence residuals, evaluated just outside the band (at $\alpha = 1.00001$, $\beta = 1e-06$, $q_{\text{loss}} = 1.00 \times 10^{-6}$) where the general formulas run and a nonzero gap is possible: those residuals are 1.66 ×10^{-5}, 1.24 ×10^{-5}, and 1.12 ×10^{-5} respectively, and any failure of those quantities to shrink toward the limit would falsify the containment claim. The posterior row is a measured identity on the general code path, so its near-zero residual is itself the falsification surface. The aggregate row's zero at exactly $c = 0$ is likewise branch-exact; the identity is additionally exercised on the iterative code path at near-zero robustness, where the consensus must still land on the log-linear pool to tight tolerance.

Table 2: Recovery residuals: the largest observed discrepancy between each robust primitive and its standard-Bayes limit, over the recovery band. Each is a maximum absolute difference in the natural units of the quantity (pmf entries for the posterior and aggregate rows, nats for the divergence and loss rows); the aggregate, divergence, and loss rows are exactly zero (bit-identical) and the posterior row is exact to machine precision (about one ULP), so the limits are verified identities, not approximations. The labeled presentation of these residuals lives in sec. 7.1.

Identity (owner statement)	Trusting limit	Tested residual
Rényi → KL (Lemma 3, eq. 3)	$\alpha \to 1$	0
rcce → NLL (Proposition 4, eq. 5)	$q_{\text{loss}} \to 0$	0
β-loss → NLL (Proposition 4, eq. 4)	$\beta \to 0$	0
generalized posterior → Bayes (Corollary 6, eq. 6)	KL, NLL	5.55e-17
`robust_aggregate` → log-linear pool (Theorem 5, eq. 7)	$c = 0$	0

The central project identity is the aggregation collapse of eq. 7: `robust_aggregate(robustness=0)` equals the log-linear pool eq. 6. Its source bridge is deliberately narrow. Take a finite common support with $q_n(s) > 0$ and represent each Eq. 7 softmax input as a posterior log potential $m_n(s) = \log q_n(s) + \kappa_n$, with κ_n constant in s and fixed declared weights w_n independent of the emerging consensus. Softmax then cancels the additive constants and yields the project pool. This identifies only the source equation's message-combination term; it does not identify source message construction, cavity/exclusion policy, scheduling, generative factors, or the complete protocol. Theorem 5 (sec. 5.8) states that specialization and the local $c = 0$ identity; the residual 0 above pins the latter. Corollary 6 establishes the separate client result: `generalized_posterior(KLD, NLL)` reproduces the closed-form prior-times-likelihood Bayes posterior of eq. 6 to residual 5.55e-17. Pooling such local posteriors has the stated categorical specialization only under the theorem's assumptions. The honesty contract binds here: the theorem and corollary cover only the recovery identity and the per-agent rigorous axis (Proposition 4); no statement transfers the bounded-influence guarantee to the server-side divergence-reweighting heuristic, whose positive property is the `robustness` $= 0$ limit of eq. 7. A scoped no-go rejects a declared separable objective class without certifying another.

6.2 Expected-free-energy identity as an algebraic check

The active-inference substrate that drives the studies of sec. 7 is a categorical specialization of the expected-free-energy algebra discussed by Friston et al. [Friston et al., 2024]. It decomposes the expected free energy of a policy π into two equivalent two-term forms. The risk-plus-ambiguity (cost) view and the negated pragmatic-plus-epistemic (value) view are the same scalar $G(\pi)$ rearranged [Da Costa et al., 2020, Friston et al., 2024]:

$$G(\pi) \;=\; \underbrace{\text{risk} + \text{ambiguity}}_{\text{cost view}} \;=\; -\big(\underbrace{\text{pragmatic} + \text{epistemic}}_{\text{value view}}\big). \tag{18}$$

The two views are not approximations of one another; they are the same scalar rearranged. In the implementation the shared entropy term enters both sides of the rearrangement, so the identity residual is zero by construction — it is a definitional consistency check on the decomposition's bookkeeping, not an independent measurement. The scientific content lives in the per-term semantics, which are pinned independently of the identity (see the closing clause of the proposition below):

$$(\text{risk} + \text{ambiguity}) \;+\; (\text{pragmatic} + \text{epistemic}) \;\equiv\; 0. \tag{19}$$

Proposition 7 (Expected-free-energy decomposition identity)**.** *For the categorical generative model of* `expected_free_energy.py`*, the cost decomposition and the negated value decomposition of (18) yield the same $G(pi)$, so the identity (19) holds with a residual at machine precision. The risk term is* $\mathrm{KL}(q(o|pi)\,\|\,p_C(o))$*, the ambiguity term is the expected likelihood entropy* $\mathbb{E}_{q(s)}[H[p(o|s)]]$*, the pragmatic value is the expected log-preference* $\mathbb{E}_{q_{pi}(o)}[\ln p_C(o)]$*, and the epistemic value is the state-outcome mutual information* $H[q_{pi}(o)] - \mathbb{E}_{q(s)}[H[p(o|s)]]$*; the identity follows from the cross-entropy split of the risk and the entropy split of the epistemic value. The residual of the decomposition is pinned to zero at a tolerance of* 10^{-9}*, and each term's semantics is pinned independently (deterministic likelihoods give zero ambiguity; uninformative likelihoods give zero epistemic value; preference-matched predictions lower risk).*

The executed formal-specialization diagnostic of fig. 7 uses a uniform prior over the nine locations. This is intentional: the canonical sentinel-world D_0 is a point mass at the den, and under that fully resolved prior the mutual-information term is zero because there is no state uncertainty for an observation to reduce. The uncertainty-bearing diagnostic makes the epistemic term visible without changing the canonical D_0 used by the inference and recovery studies. Thus a near-zero epistemic value is a meaningful null condition, not a missing term. fig. 7 shows the additive risk-plus-ambiguity view beside a signed pragmatic/epistemic waterfall whose terminal endpoint is $G(\pi)$, labels epistemic value as $I(s; o \mid \pi)$, and annotates the identity residual; it visualizes Proposition 7, not a fitted result, so it carries no error bars.

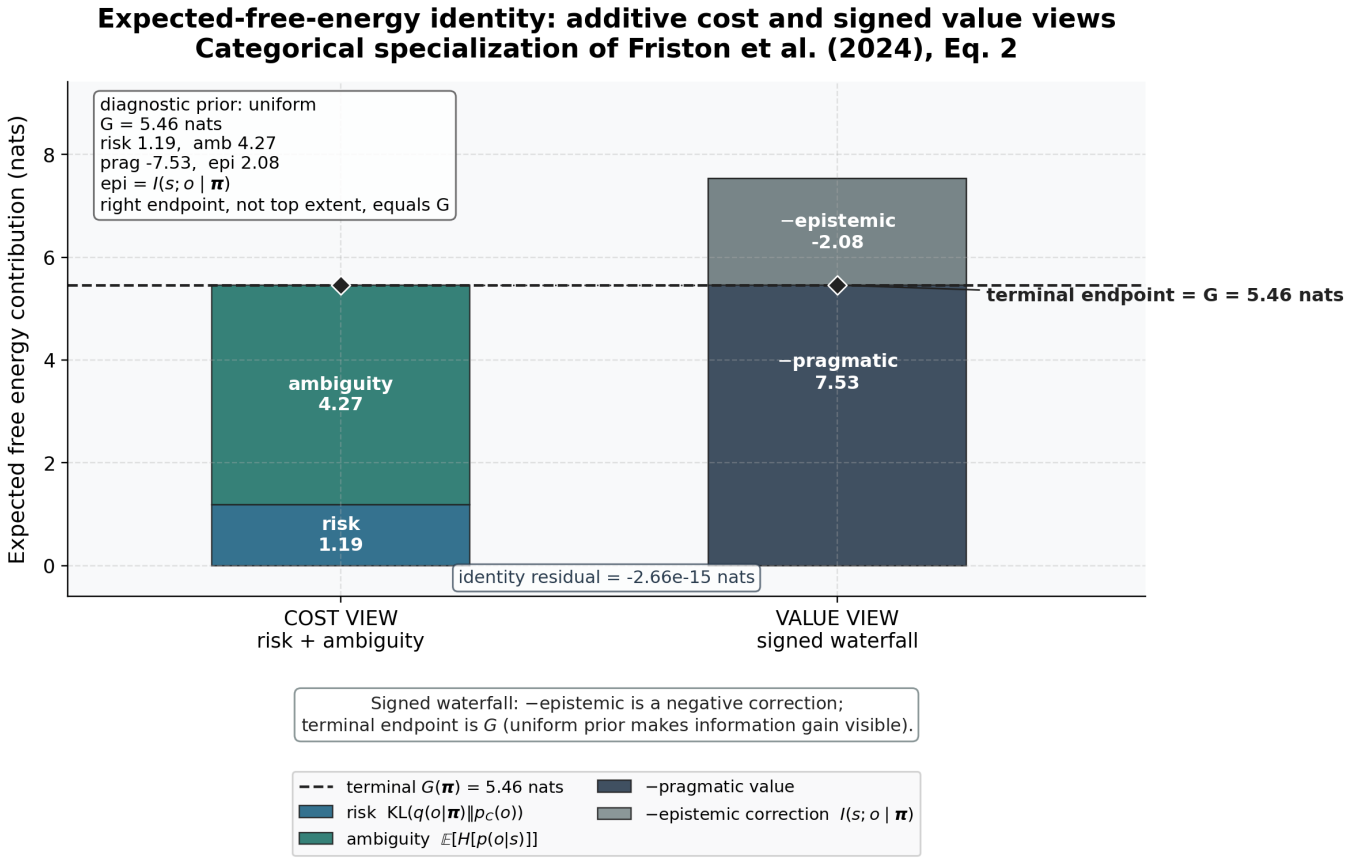


Figure 7: Expected-free-energy decomposition for the categorical generative model (expected_free_energy.py). Source relation: formal specialization of Friston et al. (2024), Eq. 2; estimand: categorical EFE identity in nats. x-axis: two views of the same identity (left, additive cost view: risk + ambiguity; right, signed value waterfall: positive minus-pragmatic contribution followed by a negative epistemic correction). y-axis: EFE contribution in nats. The heavy endpoint marker and connector, rather than the intermediate top extent, identify the terminal $G(\pi)$ value. The epistemic term is state–outcome mutual information $I(s; o \mid \pi)$; it is visible because the diagnostic prior is uniform, whereas the canonical point-mass D_0 is the corresponding zero-information null. The finite terms satisfy the identity at machine precision. This deterministic algebraic check has no error bars or independent sample size, and it does not reproduce every parameter-learning term in the source equation.

The expected-free-energy identity of eq. 19 is the action-selection counterpart of the inference-side recovery limits collected above: both are exact, closed-form, machine-checkable identities over the same categorical generative model. Together they establish that the FedGVI-federated active-inference colony of this work is built on verified algebra throughout — the per-agent generalized-Bayes update recovers standard Bayes, the aggregation recovers the project log-linear pool under its qualified categorical bridge, and the policy scoring decomposes exactly — so every robustness result in sec. 7 is a controlled departure from a known, tested fixed point rather than an unmoored claim.

6.3 Tempered aggregation free energy and the accuracy-guarantee trade

The recovery limits and the expected-free-energy identity fix the *endpoints* of the aggregation family; the remaining formal question is what a controlled departure from the unit-entropy server buys and what it costs. The objective-backed variational aggregator of sec. 11 holds its consensus-entropy term at unit weight. Freeing that single coefficient produces a one-parameter family whose only moving part is the sharpness of the consensus, and whose raw effective-weight bound is provably untouched. The following proposition isolates exactly that separation — algebra that moves versus

algebra that does not — before the interpretation subsections turn to the empirical accuracy question the algebra cannot settle on its own.

Proposition 8 (Tempered aggregation free energy)**.** *Let lambda > 0 be an entropy weight and $F_l$ambda the objective of equation 25. For a fixed effective-weight vector a, the q-block minimizer of $F_l$ambda is the tempered-softmax update in equation 26. At lambda $= 1.0$ the objective, both block updates, and the endpoint-selection rule reduce to the standard variational aggregate (Definition 10, Section 11) bit-for-bit. The a-block update and its raw effective-weight bound $a_n \leq w_n$ contain no lambda and are therefore unchanged for every lambda > 0.*

The proposition is intentionally more specific than the phrase "temperature improves robustness." It identifies exactly which part of the variational server changes when the entropy coefficient changes, and it separates that algebra from the empirical accuracy question. The objective and its update rules are a generalized-Bayes construction in the sense of [Bissiri et al., 2016, Knoblauch et al., 2022], while the particular client/server decomposition is the one implemented and tested here.

6.3.1 What the entropy weight controls

For a fixed effective-weight vector a and $\lambda > 0$, the q-block in eq. 26 is a weighted geometric pool of the local posteriors with inverse temperature $1/\lambda$. Lower λ concentrates more sharply on states that receive consistent log-belief support; larger λ spreads probability mass and retains more entropy. As $\lambda \downarrow 0$, the positive-temperature expression approaches a winner-take-most consensus (subject to ties and finite numerical support), whereas large λ approaches a flatter distribution. The implementation exposes that endpoint as a separately defined deterministic tied-argmax rule; it does not substitute $\lambda = 0$ into the objective or coordinate update. This is a controlled change in the consensus geometry, not an automatic outlier detector.

The coupling matters. Although the formula for the a-block does not contain λ, the fixed point can still change because a_n is evaluated at the new q. The correct statement is therefore conditional: for any current consensus, the effective-weight update and the raw bound $a_n \leq w_n$ are unchanged; after alternating updates, different temperatures can reach different coupled (q, a) fixed points. This distinction prevents the temperature result from being read as a theorem that the normalized influence or accuracy is invariant in λ.

6.3.2 Recovery at the qualified log-linear-pool corner

At the configured default $\lambda = 1.0$, the entropy coefficient is the unit coefficient used by the original variational aggregator. The implementation therefore recovers that aggregator bit-for-bit, including its block updates and its endpoint-selection rule. Turning the robustness strength c to zero then sets every server weight to its base value and gives the tempered log-linear pool. At the default temperature this is the ordinary log-linear pool of eq. 6. Under the shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, that is the categorical specialization of Eq. 7's message-combination term; it is not the complete belief-sharing protocol of [Friston et al., 2024]. Away from that default, the result is a tempered generalization of the pool and should not be described as Friston's Eq. 7 itself.

This nested limit is useful for interpretation. The $c \to 0$ limit identifies the aggregation family with a known consensus operator; the $\lambda = 1.0$ slice identifies the objective-backed implementation used in the main server comparison. Neither limit grants the server-side `robust_aggregate` heuristic a variational objective or a bounded-influence guarantee. The FedGVI literature's rate and robustness results remain attached to their stated loss, divergence, and sampling assumptions [Mildner et al., 2025b,a].

6.3.3 What the accuracy–guarantee trade can establish

The executed grid over $\lambda \in \{0.1, 0.2, 0.3, 0.5, 0.7, 1\}$ is a finite sensitivity study, not a search over a continuous optimum. It asks whether any tested temperature narrows the point-accuracy gap to the sharp heuristic while retaining the same effective-weight update. The closest tested temperature is $\lambda^* = 0.3$, with observed gap 0.0008. These tokens are computed from the executed contaminated-colony trials and are reported with the grid definition so a reader can reproduce the selection rule.

The result has two distinct readings. If a lower temperature improves the paired point-accuracy comparison, it provides design evidence that entropy regularization can be tuned rather than accepted as a fixed conservatism penalty. If no tested temperature closes the gap, that negative result is still informative: within this objective family, the same entropy mechanism that keeps consensus diffuse can limit exact point recovery under confident contamination. In either case, the grid does not identify a universally best temperature, establish minimax robustness, or transfer the variational raw weight bound to a different estimator. Generalized-Bayes calibration and robust-loss theory motivate the family, but only the executed categorical design supports the present finite-grid statement [Bissiri et al., 2016, Knoblauch et al., 2022, Mildner et al., 2025b].

6.3.4 Publication-facing interpretation

The practical contract is consequently three-part. Use the default temperature when exact compatibility with the tested variational server is the priority. Explore the declared grid when the application can trade concentration against the same raw effective-weight control. Treat any selected temperature as a configuration-specific empirical choice until it is tested under a new contamination mechanism, colony size, sensor model, or loss. This is the appropriate bridge between the formal objective and the active-inference setting: it exposes a tunable consensus geometry while keeping recovery, guarantee, and accuracy claims on separate evidence tracks.

7 Results: recovery checks and study suite

Every quantitative assertion in this and the following results sections is a generated token, hydrated by the manuscript-variable generator from analysis outputs produced by `src/analysis/workflow.py` and `src/fedference/experiments/` — no number is transcribed by hand. The studies implement categorical source-mechanism analogues of the colony belief-sharing scenario [Friston et al., 2024] and add the contaminated-sentinel robustness sweep and the federated neural-network baseline that are this paper's robust-federated-learning contribution. All runs are deterministic under seed 0.

We lead with the recovery limits, not with a study, because they are what makes the studies a single coherent system rather than a collection of unrelated experiments. The generalized-Bayes machinery of sec. 5 contains the standard-Bayes client corner, while the server has the exact project-local zero-robustness log-linear-pool identity. Under the explicit bridge in sec. 5.8, that pool is a categorical specialization of Eq. 7's message-combination term rather than the complete source protocol. We verify those limited identities to machine precision before reporting anything built on top of them.

7.1 Recovery limits: standard-Bayes and project-pool corners are exact to machine precision

The identities that anchor every result are the client recovery of standard Bayes at the KL/NLL/$\beta \to 0$ and $q_{\text{loss}} \to 0$ limits plus the project-local server recovery to the log-linear pool at $c = 0$ — the scoped claims of sec. 6 (Corollary 6, Lemma 3, Theorem 5). These are not figures but exact equalities, pinned by the locked core test suite. Under the theorem's shared-support, posterior-log-potential, and fixed-weight assumptions, the server pool specializes Eq. 7's message-combination term; it does not reproduce the source construction in full. Robustness is a tested extension that vanishes at the stated recovery limits.

The five residuals below are the maximum absolute deviations between each generalized-Bayes object and the standard object it must reproduce in the trusting limit. Each is a deterministic constant of the mathematics, not a per-run sample: it is reported as the maximum absolute deviation over the recovery band, which is exactly 0 where the implementation evaluates the closed form at the limit (the Rényi/loss switch) and otherwise a tiny floating-point residual:

- The server-side aggregator at zero robustness equals the log-linear pool (eq. 7, Theorem 5): maximum absolute deviation 0. This is the *naive-recovery* limit of the server-side heuristic — the only property proven for that axis (see sec. 3.3).
- The generalized posterior under the KL divergence and the NLL loss equals the closed-form prior×likelihood Bayes posterior (eq. 6, Corollary 6): maximum absolute deviation 5.55e-17.
- The Rényi divergence recovers KL as $\alpha \to 1$ (eq. 3, Lemma 3): residual 0.
- The β-loss recovers the NLL as $\beta \to 0$ (eq. 4, Proposition 4): residual 0; and the robust categorical cross-entropy recovers the NLL as $q_{\text{loss}} \to 0$ (eq. 5, Proposition 4): residual 0.

Because the Rényi divergence and the two categorical losses switch to their exact closed form inside narrow numerical-stability bands around the limit point (the Rényi switch band for α and the categorical-loss switch band for q_{loss} and β), the three zero residuals above confirm that branch equals the standard object — not, by themselves, that the *general* formula converges there. As a genuine (non-branch) convergence witness, evaluating each general formula strictly *outside* its switch band — $q_{\text{loss}} = \beta = 1.00 \times 10^{-6}$ for the categorical losses and $\alpha = 1.00001$ for the Rényi divergence — gives residuals 1.12e-05 (rcce), 1.24e-05 (β-loss), and 1.66e-05 (Rényi): nonzero (a small multiple of the input offset itself, as the first-order Taylor behavior near the limit predicts) yet still several orders of magnitude below the $O(1)$ scale of the loss/divergence values being compared, and shrinking monotonically as the offset shrinks toward the switch band (verified in `tests/fedference/test_core_identities.py`). This is evidence that the general formula itself converges to the standard-Bayes limit, not merely that the implementation switches to it exactly at the corner.

The first residual is the naive-aggregate limit of the *server-side* heuristic (Theorem 5); the latter four are the per-agent generalized-Bayes recoveries (Corollary 6 + Proposition 4) and the divergence-family recovery in the Rényi limit (Lemma 3, eq. 3) that define the theorem-bearing FedGVI axis under matching assumptions. Keeping the three axes distinct at the level of the recovery limits is what lets the robustness claims of sec. 7.5 and sec. 7.6 rest on the per-agent axis without leaning on the heuristic.

257 of 259 acceptance criteria are verified. The pure-NumPy/SciPy core carries project test coverage of 90.04% (gate $\geq 90\%$), with every stochastic step threaded through a single seeded `np.random.default_rng(0)`. sec. 10 records the full environment fingerprint, and the expected-free-energy identity that underwrites the active-inference substrate is proven and visualized in sec. 6.2 (fig. 7).

7.2 Belief sharing lowers free energy at the project-pool corner

With the standard-Bayes client limit and project-pool recovery identity pinned exactly (sec. 7.1), the study suite opens one step away from them. The first study's estimand is the colony's mean variational free energy under two communication conditions; the design is a categorical source-mechanism analogue of the colony belief-sharing result [Friston et al., 2024], implemented at the stated categorical recovery limits. A colony of 7 sentinels each observe the same hidden creature location through an independent noisy sensor (acuity 0.55) and form a one-step variational posterior. When the colony runs one federated belief-sharing round — the standard log-linear pool, which is the `robustness=0` corner of the aggregation identity eq. 7 proven in Theorem 5 — each agent's posterior moves toward the cross-agent consensus via the belief-sharing round eq. 10. Under the theorem's shared-support, posterior-log-potential, and fixed-weight assumptions, that pool specializes Eq. 7's message-combination term, not the complete source protocol; when held incommunicado, agents keep their private posteriors. Scoring each belief against the colony's joint evidence yields the "two heads are better than one" reduction, reported here as earned quantities rather than asserted:

- Mean variational free energy, **communicating**: $\bar{F}_{\text{share}} = 13.2190$ (across-seed 95% bootstrap CI $[12.9656, 13.4685]$ over $n = 480$ seeds). The single illustrative seed-0 run has its own colony mean 15.8115, with a per-agent 95% bootstrap CI of $[15.6177, 16.0647]$ over its $n = 7$ agents — that interval characterizes the displayed run's per-agent spread, not the across-seed mean
- Mean variational free energy, **incommunicado**: $\bar{F}_{\text{solo}} = 16.5298$
- Free-energy reduction from sharing: $\Delta\bar{F} = 3.3109$ (communicating is strictly lower)

The across-seed colony means above are computed over $n = 480$ seeds. The communicating colony also reaches higher mean true-state accuracy (0.7302) and lower mean surprise (0.3775) than its members reach alone. Sharing pulls each private posterior toward the joint minimizer, so the mean free energy when communicating sits below the incommunicado value.

fig. 8 reports the headline gap; fig. 9 shows the per-agent mechanism behind it.

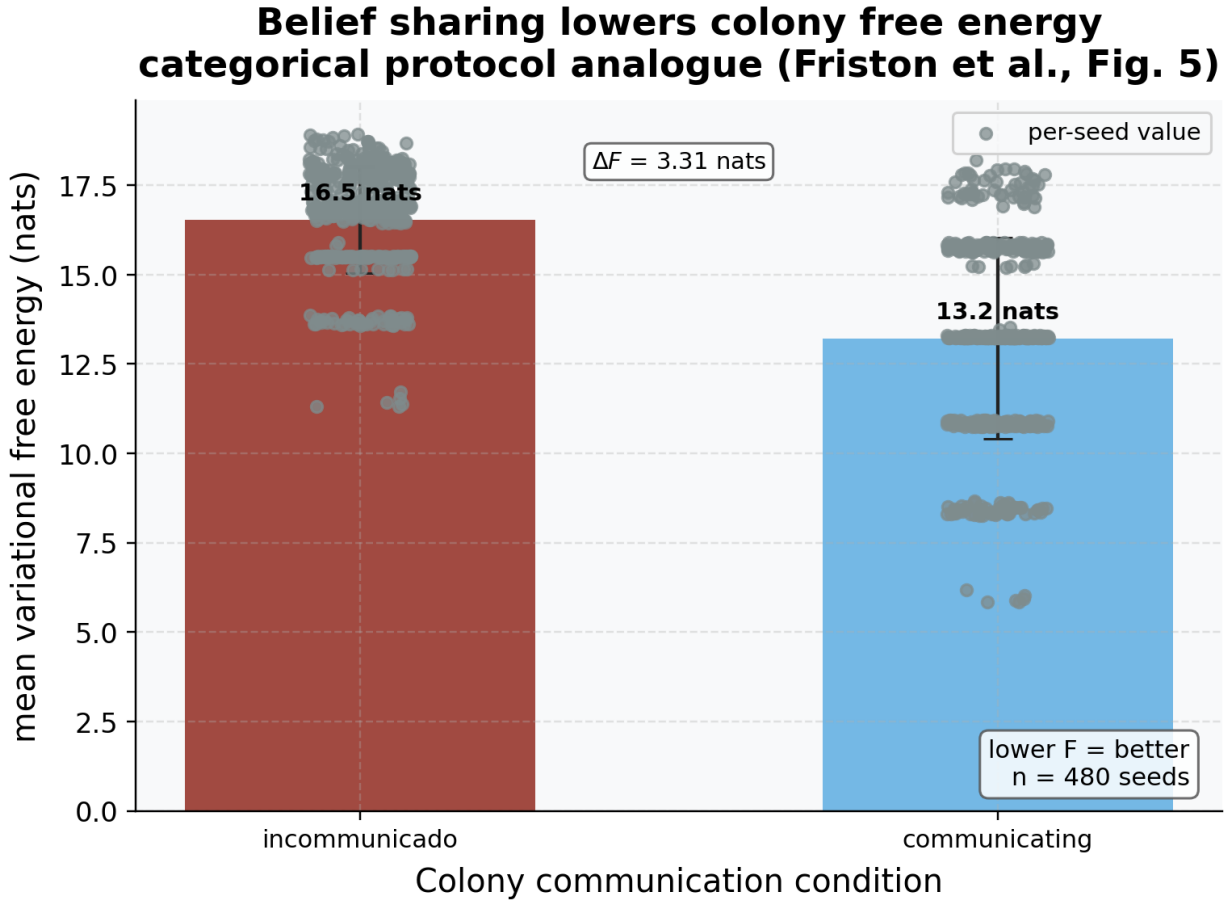


Figure 8: Mean variational free energy of the sentinel colony. Source relation: source-mechanism analogue to the belief-sharing mechanism in Friston et al. (2024), Fig. 5; estimand: colony-mean variational free energy in nats; uncertainty: across-seed spread over independent seeds. The sentinel colony (7 agents, acuity 0.55) under two communication conditions. x-axis: condition (incommunicado vs. communicating — one standard belief-sharing round), plotted in that order; y-axis: colony-mean variational free energy in nats. Bars show the mean free energy averaged across $n = 480$ independent random seeds, with whiskers marking $\pm$one across-seed standard deviation and grey points overlaying the individual per-seed values. The communicating bar is strictly lower than the incommunicado bar, with $\Delta\bar{F} = 3.3109$ nats — the quantitative "two heads are better than one" result. Each seed is a fully deterministic run; the whiskers are the across-seed spread, not a bootstrap or resampling interval.

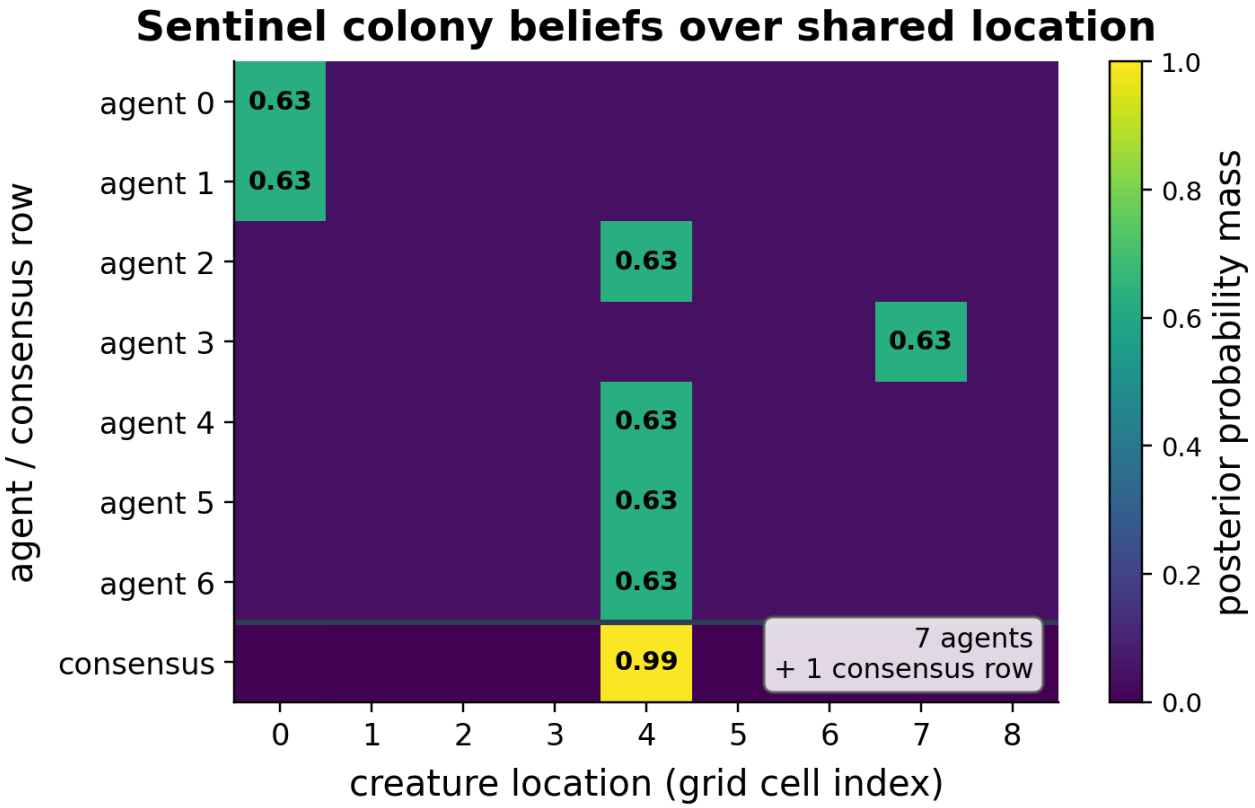


Figure 9: Single-panel belief heatmap over the hidden creature location. Source relation: original project diagnostic supporting the Study 1 analogue; estimand: posterior probability mass by hidden state; uncertainty: deterministic single-seed display. The hidden creature location (7 sentinels, acuity 0.55, seed 0). x-axis: hidden-state grid cell (creature location, 9 cells); rows: the 7 individual agents' private posteriors (one row per agent, dominant cell annotated), plus a bottom consensus row — separated by the divider line — holding the federated consensus fused from those posteriors by the cavity-exclusion round defined in the methods. Each private posterior concentrates only moderately on the cell its noisy observation suggests; the consensus row concentrates far more sharply on the true location, and that after-sharing concentration is the mechanism behind the free-energy reduction reported by the free-energy comparison. All cell values are deterministic posterior probabilities for the single displayed seed; no error band is applicable.

7.2.1 Three robustness axes remain distinct in the results

Before the contaminated studies, we fix the honesty boundary that governs every robustness claim in this paper, because the belief-sharing round above is exactly where the axes diverge once contamination is introduced. The robust extension of belief sharing lives on **three distinct axes** that must not be conflated.

The **per-agent axis** is the generalized-Bayes update each sentinel runs locally: the β-loss and robust cross-entropy clients of eq. 4 and eq. 5. This axis follows the FedGVI objective [Mildner et al., 2025b] and carries the cited bounded-influence result only under that theorem's matching assumptions. The recovery limits of sec. 7.1 show the per-agent update reduces to standard Bayes; the separately stated server identity returns the project log-linear pool.

The **server-side axis** is the `robust_aggregate` divergence-reweighting that down-weights agents at pooling time. This is a complementary *heuristic*. Its only proven property is the naive-recovery limit of Theorem 5 — at zero robustness it equals the standard log-linear pool (eq. 7) — and it carries **no** bounded-influence guarantee. No figure, table, or sentence in sec. 7.5 or sec. 7.6 grants the server-side heuristic the guarantee that belongs to the per-agent axis. The contaminated sweep that follows reports the axes side by side and labels which is which at every step.

The **variational server axis** is `variational_aggregate`. It is also server-side, but it is not the same claim as the sharp heuristic: it descends the stated aggregation free energy (eq. 8), carries the derived effective-weight bound proved in the supplement, and pays for that property with a conservative maximum-entropy bias. The contaminated sweep therefore reports behavior for the sharp heuristic and cites the variational rule only where the objective-backed guarantee is actually in force.

7.3 Language acquisition follows conjugate Dirichlet updating

Where the first study fused fixed beliefs in a single round, the second asks whether a single agent can *acquire* the likelihood of its shared world at all, and how quickly it does so. The design is a categorical source-mechanism analogue of the language-acquisition mechanism discussed by Friston et al. [Friston et al., 2024]: each configured seed runs an agent that learns the likelihood of its shared world by conjugate Dirichlet updates over 24 count batches, the count update eq. 14. The recorded trajectory is KL(true $A \,\|\, \text{learned } A$) before each batch — it starts at the flat-prior maximum and declines monotonically toward zero as the agent "acquires the language" of its world. The KL here is the same divergence whose $\alpha \to 1$ Rényi limit is established by Lemma 3 (eq. 3), so the learning curve and the recovery limits measure the same object.

- Initial KL (flat prior): 3.4231
- Final KL (after 24 batches): 0.0027
- Total KL reduction: 3.4204
- Trajectory points: 25 ordered learning steps per seed
- Pointwise seed bootstrap: 95% intervals over $n = 480$ independent seeds
- Trajectory monotone-decreasing: Yes

The monotone decline to a final KL of 0.0027 is the demonstrated quantity behind the acquisition claim: under the tested count schedule, the learned likelihood moves toward the true generative likelihood as conjugate counts accumulate. This finite trajectory is evidence of the update's behavior, not a convergence-rate result for arbitrary data-generating processes.

fig. 10 plots the full learning curve and its CI band.

7.4 Bayesian model reduction selects supported structure

The first two studies fixed the model's structure and asked how well its beliefs and learned parameters track the world; the third asks whether the model can also shed structure it never needed. The estimand is a Bayesian-model-reduction free-energy difference; the design is a categorical BMR diagnostic related to the structure-emergence mechanism discussed by Friston et al. [Friston et al., 2024], through the Bayesian-model-reduction lineage [Friston and Penny, 2011]. A full Dirichlet model carries a redundant state — one column the data never support — ranging over $n = 4$ candidate states. Bayesian model reduction scores swapping the prior for a *reduced* prior that prunes that column; the free-energy difference ΔF is the model-reduction objective eq. 16. This is a single deterministic evidence comparison, so there is no resampled sample and, by design, no confidence interval or paired test. The structure-learning frame here is the discrete-state model-selection thread the active-inference community has developed [Smith et al., 2022], applied to a colony's shared generative model.

ΔF is positive for the correct (redundant) pruning — the simpler model has more evidence and the run converges on it — and negative for the control pruning of a well-supported column, which is correctly rejected:

- ΔF, pruning the **redundant** column: 3.68 (positive — reduction accepted)
- ΔF, pruning a **supported** column (control): -27.67 (negative — reduction rejected)
- Emergence converged (redundant accepted, supported rejected): Yes

The sign pattern $\Delta F_{\text{redundant}} > 0 > \Delta F_{\text{supported}}$ is the demonstrated emergence verdict: the colony's generative model prunes the structure its data never support and retains the structure they do.

fig. 11 contrasts the two prunings and annotates the convergence verdict.

Studies 1–3 all ran in a *trusting* world, where every broadcast belief is honest. The contamination sweep that follows removes that assumption, and it is the point at which the three robustness axes of sec. 7.2.1 begin to diverge.

7.5 Contamination sweep: regime-dependent server behavior under declared attacks

This contamination experiment compares an active-inference colony with server operating points named for the FedGVI client-divergence vocabulary [Mildner et al., 2025b]; the cited sources do not make this categorical belief-fusion comparison. A colony of 7 sentinels broadcasts soft beliefs about a shared state, while 2 saboteurs mix toward a confident-wrong delta at each contamination `rate`. `KLD` is the standard log-linear pool (server robustness

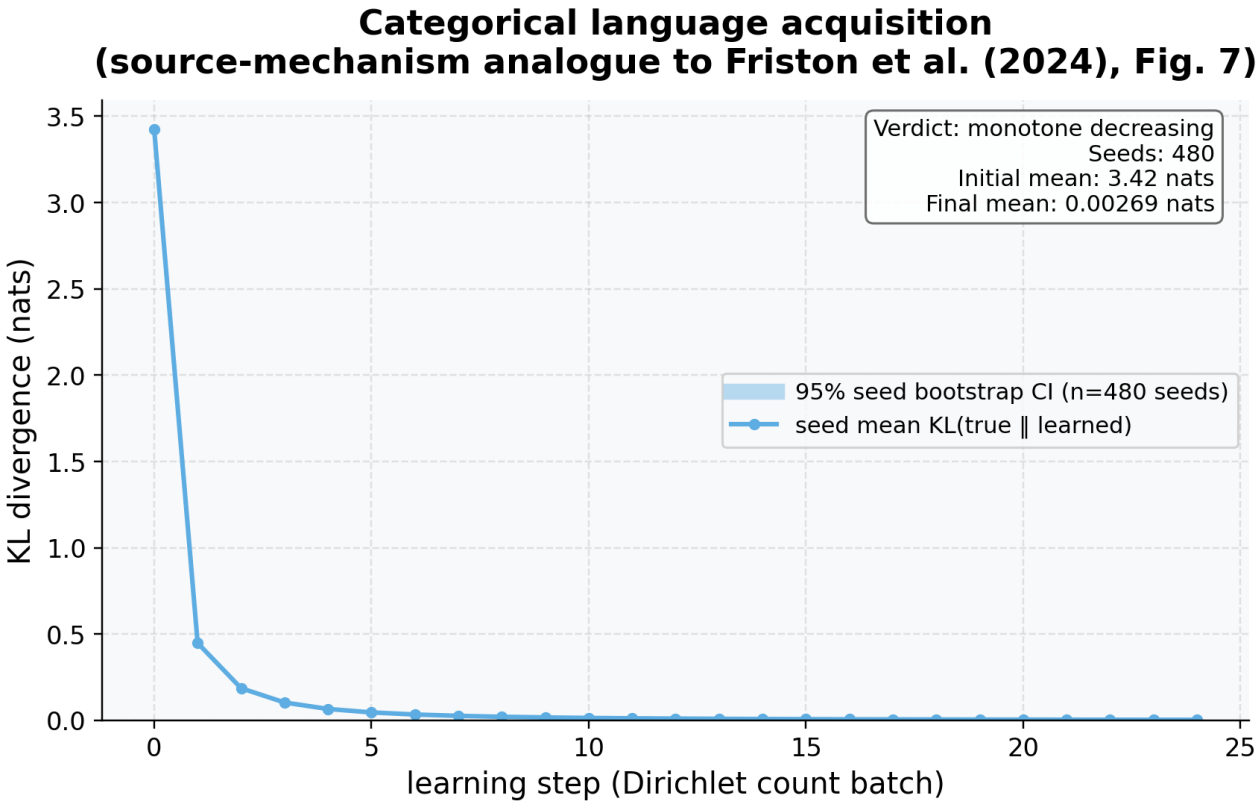


Figure 10: Seed-mean KL divergence from the true likelihood A to the learned likelihood A. The plotted quantity is KL(true $A \,\|\, \text{learned } A$). Source relation: source-mechanism analogue to Friston et al. (2024), Fig. 7; estimand: seed-mean KL in nats by ordered learning step. x-axis: ordered Dirichlet count batch, from the flat prior at zero through all 24 batches (25 points per seed); y-axis: summed per-column KL divergence between the true likelihood and the current expected likelihood, in nats. The solid line is the mean over 480 independent configured seeds, and the shaded band is the pointwise 95% percentile-bootstrap interval resampling seeds at each learning step. The replication unit is seed, not the ordered trajectory points. The mean curve falls monotonically from 3.4231 nats to 0.0027 nats (total reduction 3.4204 nats, computed from the unrounded endpoints); the computed monotone-decreasing verdict is Yes. This reduced categorical protocol is related to, but does not exactly reproduce, the richer multi-episode protocol in Friston et al. (2024).

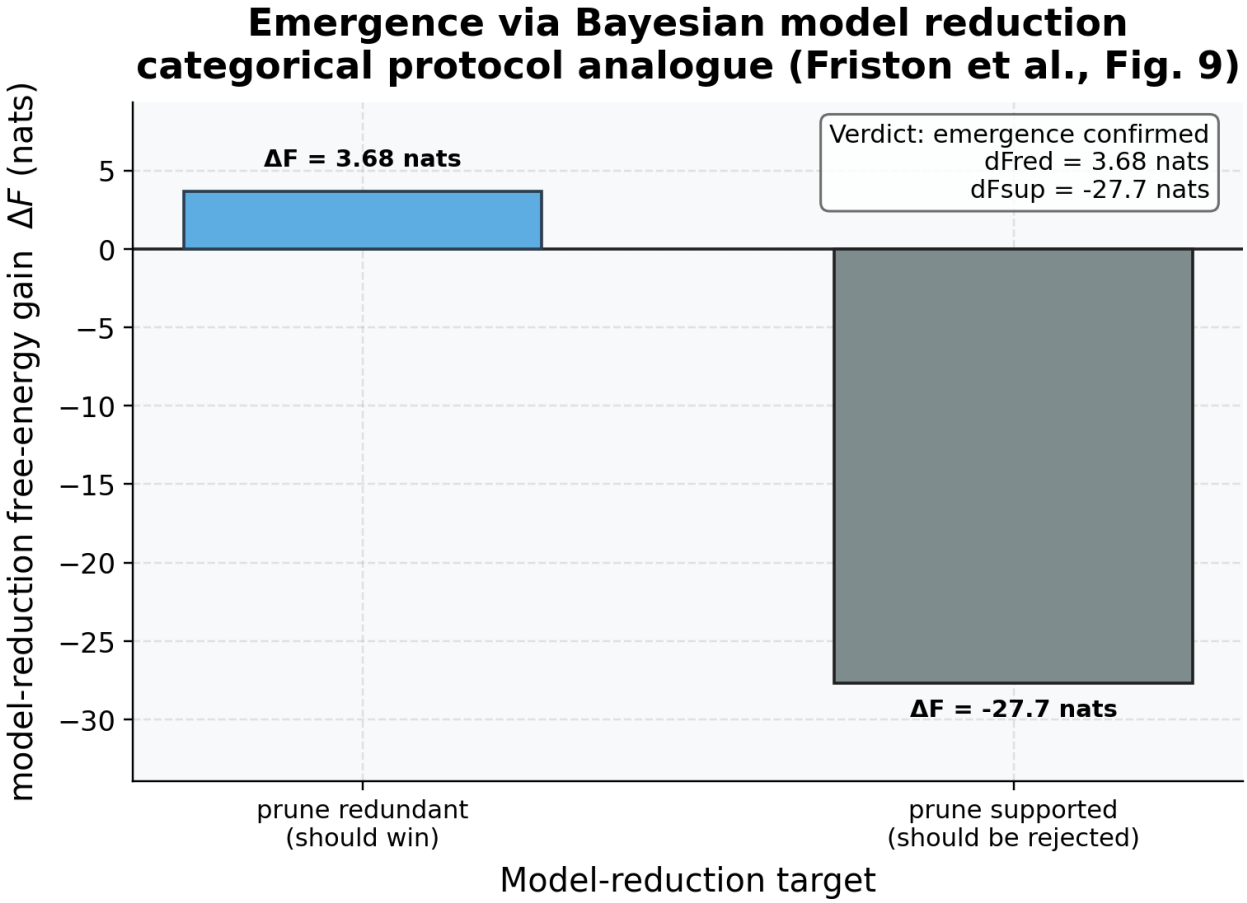


Figure 11: Bayesian-model-reduction (BMR) free-energy difference. Source relation: source-mechanism analogue to the model-reduction mechanism in Friston et al. (2024), Fig. 9; estimand: BMR ΔF in nats; uncertainty: deterministic closed-form comparison. ΔF for two candidate likelihood-column prunings in a colony with $n = 4$ hidden states. x-axis: the candidate pruning (redundant column vs. supported-column control); y-axis: ΔF in nats, where a positive value means the reduced model has more evidence and the pruning is accepted. The redundant-column bar is positive ($\Delta F = 3.68$ nats) — the data never supported this structure, so pruning it is the correct decision — while the supported-column control bar is negative ($\Delta F = -27.67$ nats), correctly rejected. The opposing signs constitute the emergence verdict (Yes). No error bar applies: BMR is a deterministic closed-form comparison on a single posterior.

0, eq. 7); the other labels denote `robust_aggregate` heuristic settings with constants KLD (c=0.00), RKL (c=1.50), AR (c=1.30), beta (c=1.70), rcce (c=1.60), not literal client losses or divergences. The estimand is consensus mass q(true state).

As the contamination rate rises across $\{0, 0.225, 0.45, 0.675, 0.9\}$, the **standard** (`KLD`) consensus accuracy degrades monotonically:

Table 3: Consensus accuracy q(true state) by contamination rate and configured server operating point ($n = 1$ deterministic sweep per cell). `KLD` is the standard log-linear pool (eq. 6); the other columns use the fixed heuristic constants listed above in the same seeded colony. As the saboteurs capture more belief mass, `KLD` falls monotonically while at least one robust member remains above the stated accuracy threshold. This table is descriptive; inferential paired evidence appears below.

Contamination rate	KLD	RKL	AR	beta	rcce
0	1.000	0.997	0.998	0.996	0.996
0.225	0.999	0.993	0.995	0.990	0.992
0.45	0.995	0.987	0.990	0.984	0.986
0.675	0.975	0.985	0.989	0.981	0.983
0.9	0.693	0.984	0.988	0.980	0.982

- Standard accuracy degrades monotonically with rate: Yes
- At the worst rate (0.900), at least one robust member stays at or above the accuracy threshold 0.50: Yes (standard accuracy there is 0.6928, highest pooled robust mean 0.9880; worst-rate display method beta)

The rate trend above is a single deterministic sweep per cell. To attach a per-rate paired test, each contamination rate is re-run as $n_{\text{trial}} = 960$ matched trial replicates nested within the fixed seeded world, and every robust member is compared against the standard pool at that rate; the resulting p-values are BH-deflated per method ([Benjamini and Hochberg, 1995]; $\alpha = 0.05$). This table is a rate-resolved diagnostic, not a continuous-family proof: only the displayed method-rate cells with `Reject = Yes` survive their own per-method BH family:

Table 4: Per-contamination-rate standard-vs-robust paired tests (matched-pairs Wilcoxon, [Wilcoxon, 1945, Fay and Proschan, 2010], $n_{\text{trial}} = 960$ matched trial replicates per cell), BH-deflated within each method's rate family. The trial-level result is conditional on the fixed seeded world and is not a claim about 960 independent worlds. The `KLD` baseline is excluded because it is the standard reference, not a self-contrast. The displayed d-equivalent is a rank-biserial-derived transform, not raw Cohen's d; the signed-saturation marker flags contrasts where the rank-biserial correlation saturates at ± 1. These contrasts decorate the server-side `robust_aggregate` heuristic only.

Method	Rate	Rank-biserial-derived d-equivalent	Label	Raw p	q	Reject
RKL	0	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
RKL	0.225	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
RKL	0.45	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
RKL	0.675	13.79	large	1.88e-155	2.35e-155	Yes
RKL	0.9	-0.37	small	1.06e-06	1.06e-06	Yes
AR	0	saturated (r=-1)	large	1.11e-158	1.47e-158	Yes
AR	0.225	saturated (r=-1)	large	1.11e-158	1.47e-158	Yes
AR	0.45	-303.73	large	1.13e-158	1.47e-158	Yes
AR	0.675	160.07	large	1.17e-158	1.47e-158	Yes
AR	0.9	-1.57	large	1.62e-61	1.62e-61	Yes
beta	0	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
beta	0.225	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
beta	0.45	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
beta	0.675	2.81	large	6.29e-106	7.86e-106	Yes
beta	0.9	0.61	medium	6.04e-15	6.04e-15	Yes
rcce	0	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
rcce	0.225	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
rcce	0.45	saturated (r=-1)	large	1.11e-158	1.85e-158	Yes
rcce	0.675	5.67	large	2.37e-141	2.96e-141	Yes
rcce	0.9	0.14	negligible	6.54e-02	6.54e-02	No

7.5.1 Earned robustness verdict at the decisive rate

The headline "robust beats standard" claim is *computed*, never asserted. Across 960 matched trial replicates nested within the fixed seeded world at contamination rate 0.800 — each redrawing the heterogeneous healthy colony while holding the run's true state and attack target fixed — every robust divergence's per-trial consensus accuracy is compared against the standard pool's by the matched-pairs Wilcoxon signed-rank test ([Wilcoxon, 1945]), and the family of p-values is deflated with Benjamini–Hochberg FDR ([Benjamini and Hochberg, 1995]; $\alpha = 0.05$). A method wins if and only if BH rejects its null and its effect size is positive.

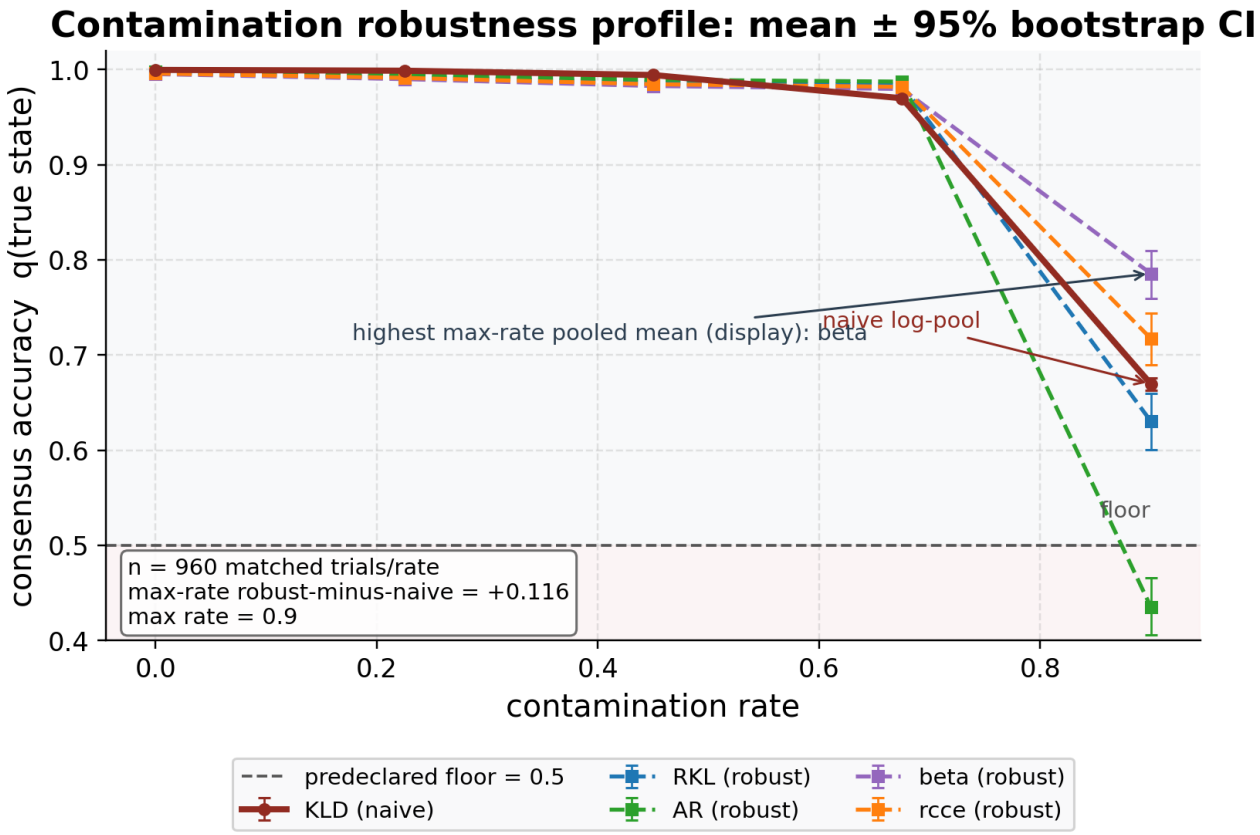


Figure 12: Consensus accuracy: probability mass assigned to the true hidden state. The plotted estimand is q(true state). Source relation: original project robustness extension; estimand: true-state probability mass; uncertainty: matched-trial percentile-bootstrap intervals over configured trials. The colony of 7 sentinel agents (acuity 0.55, of which 2 are saboteurs) as a function of contamination rate. x-axis: saboteur convex-mix contamination rate, sampled over $\{0, 0.225, 0.45, 0.675, 0.9\}$; y-axis: consensus accuracy q(true state) (probability mass on the true hidden state). One curve per configured server operating point: standard `KLD` plus the robust `robust_aggregate` settings $KLD(c = 0.00), RKL(c = 1.50), AR(c = 1.30), beta(c = 1.70), rcce(c = 1.60)$; these curves do not apply the named client losses or divergences. The dashed floor is the predeclared accuracy threshold 0.50; the in-figure box reports the matched-trial sample size and the largest-rate pooled robust-minus-naive separation, where the standard `KLD` log-linear pool reaches 0.6697 and the highest pooled robust mean reaches 0.7857. Robust means are similar to or slightly below the standard pool at low contamination and some pooled robust operating points separate in favor of the robust family under severe contamination; individual robust members can still fall below the floor at the largest rate. The linear y-axis is deliberately truncated just below the threshold band so the curves, floor, and CIs remain legible. The plotted curve is the matched-trial mean over 960 trials per rate with percentile-bootstrap 95% CIs; intervals are conditional on the fixed seeded true state and attack geometry, not alternate world models. The single-colony mechanistic table above remains descriptive and deterministic. The formal verdict-rate statistical test is reported immediately below.

Table 5: Per-method paired verdict against the standard pool at verdict rate 0.800 ($n_{\text{trial}} = 960$ matched trial replicates, signed-rank test under the paired-difference assumptions of sec. 5.28). This is a fixed-world conditional diagnostic; seed-level inference is reported separately in sec. 12.2.4. Effect size is the matched-pairs rank-biserial correlation; positive values mean the robust member exceeds the standard pool. A `Wins` value is true only when the BH-deflated null is rejected *and* the effect is positive, with FDR scoped to this verdict family.

Robust divergence	Effect size	Raw p	q	Wins
AR	1.0000	1.11e-158	1.11e-158	Yes
RKL	1.0000	1.11e-158	1.11e-158	Yes
beta	1.0000	1.11e-158	1.11e-158	Yes
rcce	1.0000	1.11e-158	1.11e-158	Yes

Table 6: Standardized-effect verdict at rate 0.800 ($n_{\text{trial}} = 960$ matched trial replicates): the rank-biserial effect and its derived d-equivalent display label, the mean robust-minus-standard accuracy difference with its 95% bootstrap CI, raw and BH-deflated p-values, the observed-effect design power of the paired Wilcoxon at $\alpha = 0.05$ (alternative greater), and the prospective n_{trial} a confirmatory fixed-world replication should budget for power 0.80. The signed-saturation marker replaces a misleading literal where the rank-biserial effect saturates; the power calculation is a planning approximation, not independent evidence.

Method	Rank-biserial-derived d-equivalent	Label	Mean acc. diff	95% CI	Raw p	q	Design power	n for power	Reject
AR	saturated (r=+1)	large	0.0846	[0.0821, 0.0873]	1.11e-158	1.11e-158	1.0000	5	Yes
RKL	saturated (r=+1)	large	0.0809	[0.0783, 0.0834]	1.11e-158	1.11e-158	1.0000	5	Yes
beta	saturated (r=+1)	large	0.0764	[0.0738, 0.0790]	1.11e-158	1.11e-158	1.0000	5	Yes
rcce	saturated (r=+1)	large	0.0787	[0.0761, 0.0814]	1.11e-158	1.11e-158	1.0000	5	Yes

Table 7: Per-method consensus accuracy at the verdict rate 0.800 with 95% percentile-bootstrap CI, including the standard `KLD` baseline. The CI is conditional on the seeded matched-trial design and resamples trials, not alternate world models. The standard pool sits at 0.9021 ($[0.8993, 0.9049]$); the highest pooled robust mean among the configured members is 0.9867 ($[0.9865, 0.9869]$).

Method	n	Mean acc. @ verdict rate	95% CI
KLD	960	0.9021	[0.8993, 0.9049]
RKL	960	0.9829	[0.9827, 0.9832]
AR	960	0.9867	[0.9865, 0.9869]
beta	960	0.9785	[0.9782, 0.9787]
rcce	960	0.9808	[0.9806, 0.9810]

- Naive pool mean accuracy at the verdict rate: 0.9021 (per-trial mean over 960 trials; the bootstrap CI is in tbl. 7)
- At least one robust method is a BH-rejected positive contrast: Yes
- Headline display method: RKL (tied set: RKL, AR, beta, rcce), rank-biserial-derived d-equivalent = saturated (r=+1) (large), mean accuracy difference 0.0846 ($[0.0821, 0.0873]$), raw $p = 1.11 \times 10^{-158}$, $q = 1.11 \times 10^{-158}$, observed-effect design power 1.0000, prospective n for power 0.80 is 5
- Headline display rule (largest positive rank-biserial effect_size; stable method order tie-break; tie-break: first robust method in divergences order): observed-effect design power 1.0000 at the run's $n_{\text{trial}} = 960$; a confirmatory replication should budget $n_{\text{trial}} = 5$ for power 0.80 (prospective $n = 7$)

The standard pool degrades, the robust family has conditional contrasts, and the verdict carries paired statistics with multiple-testing deflation, bootstrap intervals, and a power analysis — all produced by the statistics module, not typed into the prose.

The headline label is a deterministic display choice, not a unique scientific winner. The complete tied set is RKL, AR, beta, rcce, the largest paired mean-difference method is AR, and the worst-rate pooled display method is beta. These may differ because they answer different descriptive questions.

7.5.1.1 Server-side heuristic axis: accurate but not guaranteed

The verdict above is reported on the **server-side heuristic axis**. fig. 13 visualizes the `robust_aggregate` divergence-reweighting that down-weights agents at pooling time. Its positive formal property is the naive-recovery limit of Theorem 5 (eq. 7, the 0 residual of sec. 7.1), and it has a scoped no-go result for a declared separable objective class, not an objective certificate. It carries *no* bounded-influence guarantee. The effect sizes, CIs, and power above decorate this heuristic's contrasts; they do not certify the per-agent generalized-Bayes guarantee, which is established separately on the per-client axis in sec. 7.6.

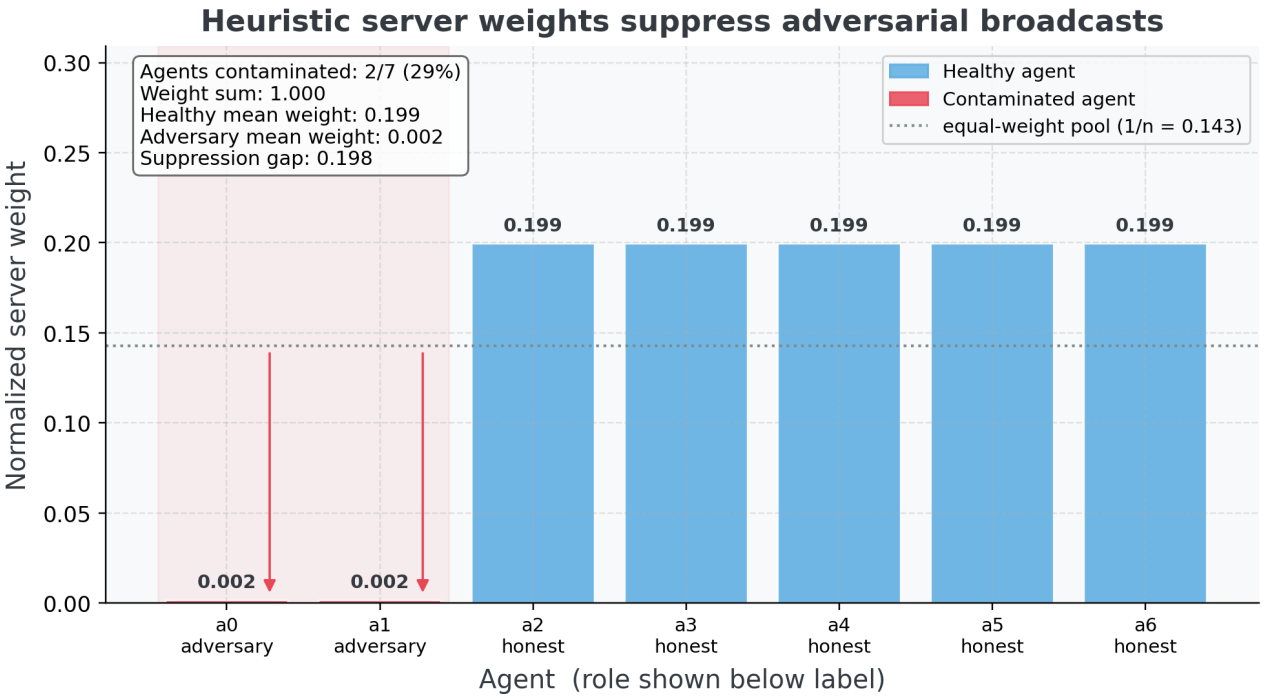


Figure 13: Server-side influence weights assigned by `robust_aggregate`. Source relation: original project server-side diagnostic; estimand: normalized pooling weight; uncertainty: deterministic single-run display. Divergence-reweighting weights for each of the 7 agents at the verdict contamination rate 0.800 (the convex-mix strength applied to each saboteur's belief — distinct from the *count* of contaminated agents, 2 of 7, reported in the in-figure box). x-axis: agent index, zero-based ($a0$ upward), with each agent's role (honest / adversary) shown beneath its label; y-axis: normalized pooling weight, with weights summing to one and the dotted reference marking the equal-weight pool ($1/n$). The 2 contaminated saboteur agents are highlighted, with downward arrows marking their suppression below the equal-weight reference. The heuristic down-weights saboteurs relative to both the equal-weight reference and the $7-2$ healthy agents. Important limitation: this is the **server-side heuristic axis only** — the reweighting is proven solely at the `robustness=0` recovery limit and does not carry the bounded-influence guarantee of the per-client FedGVI losses. Single deterministic run; no error band.

7.5.2 Variational aggregator: conservative objective-backed weight control

The heuristic axis above is the empirically sharp rule. The variational aggregator of sec. 5.8.2 (derived in full in sec. 11) is its objective-backed complement: each exact block update does not increase the stated free energy eq. 8, and a converged fixed point is coordinatewise stationary. Two diagnostics make the weight behavior concrete, both genuine `variational_aggregate` runs at robustness 1.50.

First, the descent. On a contaminated colony the free energy falls monotonically from 3.2458 to 2.3780 (a drop of 0.8678 over 11 block-coordinate iterations, converged: Yes), with a largest single-step *increase* of 8.88e-16 — machine zero, the numerical witness of the descent theorem (sec. 11.3).

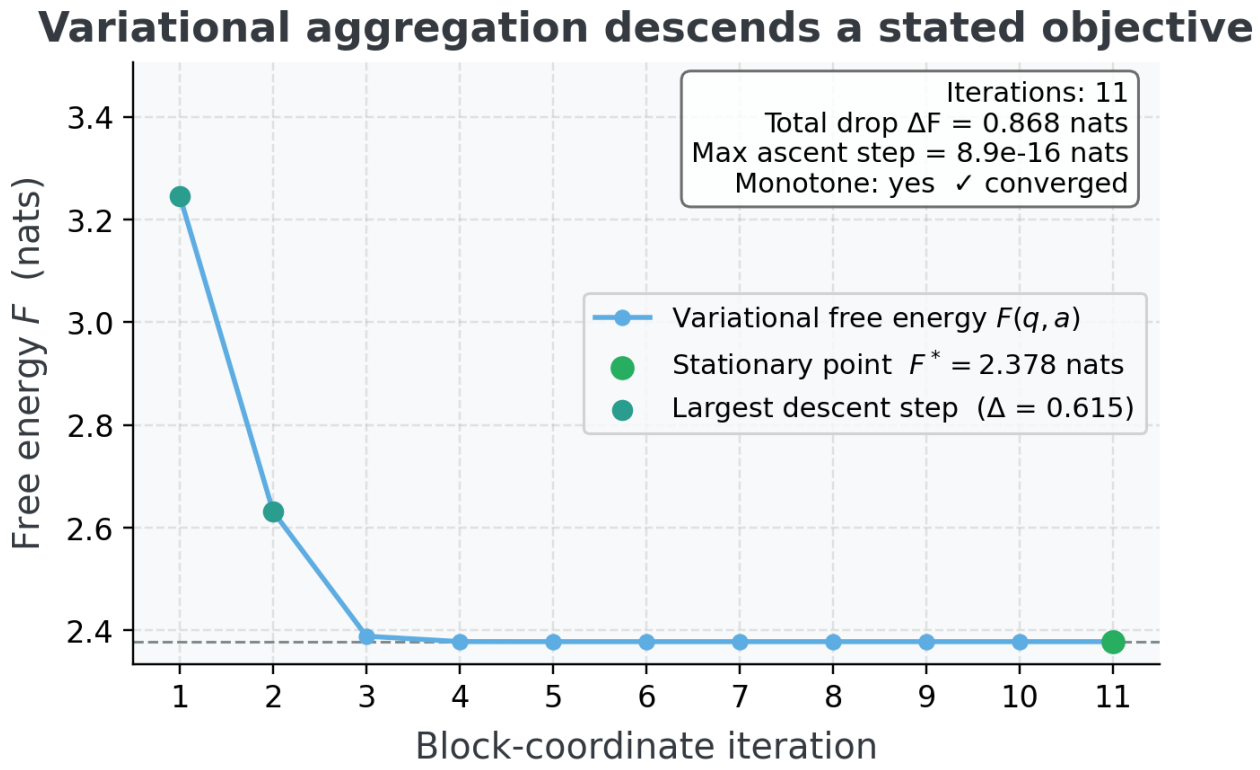


Figure 14: Variational free energy $F(q, a)$ as a function of block-coordinate descent iteration. Source relation: original project objective-descent diagnostic; estimand: free energy in nats by iteration; uncertainty: none for the deterministic seeded run. The trace is a single `variational_aggregate` fusion of a 7-agent contaminated colony (robustness 1.50). x-axis: block-coordinate iteration number; y-axis: $F(q, a)$ in nats. The curve is monotone non-increasing across all recorded iterations (largest single-step increase: 8.88×10^{-16} nats, at machine precision) and the implementation reports converged status Yes at value 2.3780 nats. This verifies objective descent on the executed run — a diagnostic the `robust_aggregate` heuristic does not provide. Deterministic seeded run; no error band.

Second, the effective-weight response. As one agent is drifted from healthy toward a confident-wrong delta, its normalized influence falls from 0.143 to below 0.001 — a factor of 267.1 below the fixed 0.143 the naive log-linear pool grants every agent regardless of how wrong it is. The gap between the falling variational curve and the flat naive line is the empirical redescending weight response, drawn.

The honest trade is conservatism: because F carries the $-H(q)$ entropy term, its consensus is the maximum-entropy distribution consistent with the weighted cross-entropies, deliberately flatter than the product-of-experts. The variational aggregator therefore does *not* win the peak-accuracy verdict of sec. 7.5.1 — that remains the sharp heuristic's role — and the two are reported as complements, never conflated: rigor-with-conservatism on one side, accuracy-without-an-objective on the other.

7.6 Client-side robustness complement: categorical FedGVI baseline

The sweep of sec. 7.5 characterizes the *server-side heuristic*. This baseline characterizes the *per-client* axis — the one that carries the provable robustness — on the setting where the robust-Bayes and federated-learning communities established their guarantees [McMahan et al., 2017, Ashman et al., 2022, Bui et al., 2018], so that the active-inference colony and the federated-learning benchmark are measured by the same robust objective.

A federated Bayesian logistic-regression colony is trained under per-client label contamination. Standard clients run the NLL / KL objective (`nll/KLD`); robust clients run the FedGVI-faithful per-agent generalized-Bayes objective with the robust cross-entropy and α-Rényi client losses (`rcce/AR`, eq. 5). This is the per-client generalized-Bayes update that recovers standard Bayes in the trusting limit (Corollary 6 + Proposition 4, the 0 and 0 residuals of sec. 7.1) and that inherits the FedGVI bounded-influence robustness [Mildner et al., 2025b]. The robust loss is the density-power / β-divergence line [Basu et al., 1998] and the generalized-cross-entropy line [Zhang and Sabuncu, 2018], folded into the generalized-Bayes objective [Bissiri et al., 2016, Knoblauch et al., 2022].

The robust client's operating point ($q = 1.00$, 200 points per client) was chosen, among the values tested, to make this margin visible rather than derived from theory; the sensitivity check below shows the qualitative result does not depend on that specific choice, which is what makes the operating point a defensible one rather than a cherry-picked one.

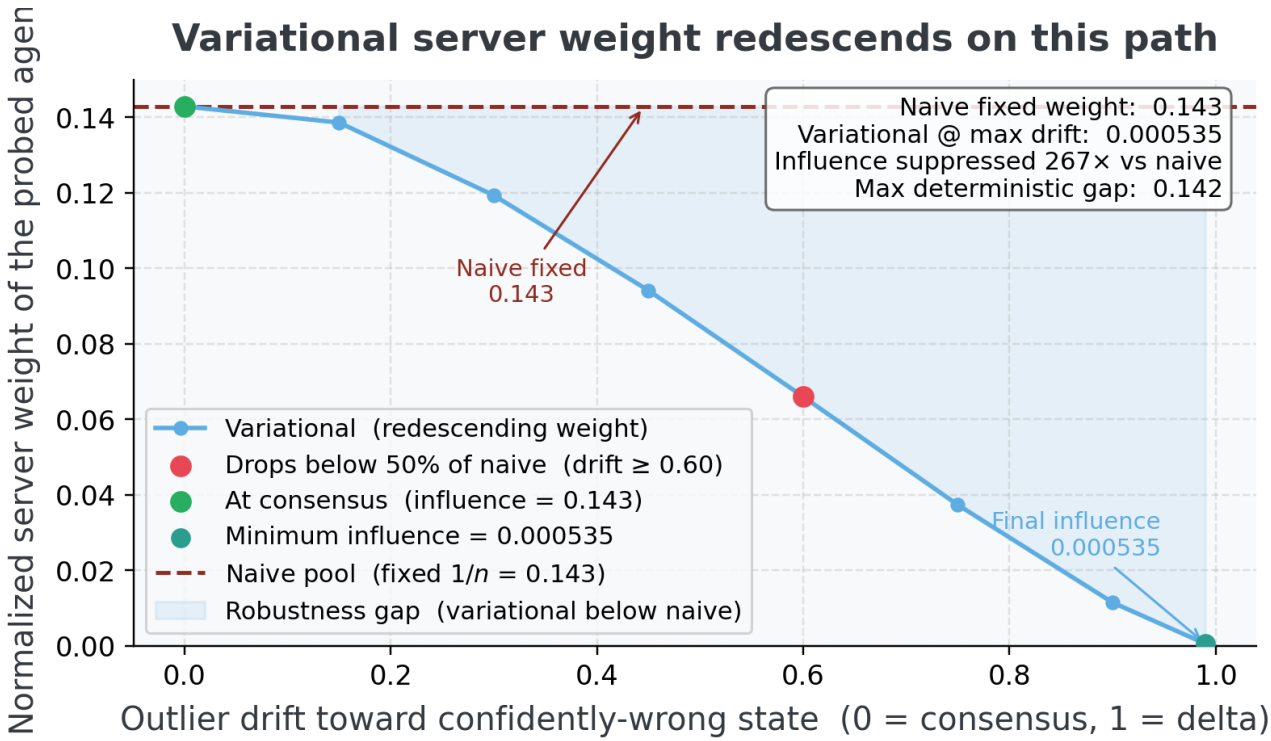


Figure 15: Normalized influence weight of one probed agent. Source relation: original project variational-server diagnostic; estimand: normalized influence weight; uncertainty: deterministic seeded sweep. The weight is shown as a function of the agent's drift toward a confident-wrong belief (delta distribution on the wrong state), under `variational_aggregate` versus the naive log-linear pool. x-axis: outlier drift — the mixing parameter carrying the probed agent's belief from the consensus posterior (zero, at consensus) to the confidently-wrong delta (one, full delta), increasing left to right; y-axis: normalized influence weight of the probed agent in the server weight vector. Under `variational_aggregate` (falling curve) the weight collapses below 0.001 as the agent goes extreme, while the naive pool holds it fixed at $1/n = 0.143$ regardless (flat line). This demonstrates redescending normalized-weight behavior on the tested path; the algebraic theorem bounds the raw effective weight, but the figure is not an estimator-level B-robustness proof. Deterministic seeded sweep over $n = 7$ agents; no error band, and no claim that the sharper `robust_aggregate` heuristic inherits this bound.

As the per-client contamination fraction rises, the robust-client curve tracks the standard curve closely at low-to-moderate contamination, then opens a genuine margin in the moderate-to-high range that peaks at 0.35 contamination (margin 0.028) — a margin that holds above a minimum threshold across a neighborhood of the robust loss parameter at more than one contamination level (`tests/fedference/test_bnn_baseline.py:: test_rcce_separation_is_not_a_knife_edge_in_loss_param`), not only at the single value plotted, and is reproducible across independent seeds rather than a single-run artifact; the plotted bands show the seed-level 95% bootstrap intervals around those means. At the most extreme 0.40 contamination level swept, both configurations decline sharply and converge again, with no reliable ordering between them; we report that point rather than omitting it, since there is no principled basis (e.g. a known breakdown point for this synthetic contamination mechanism) for excluding the one part of the sweep that does not favor the robust client.

The separation in this small logistic-regression setting is nonetheless modest and does not by itself establish a large bounded-influence effect. The recovery identities (sec. 7.1) establish implementation compatibility at the named limit; the bounded-influence result comes from the FedGVI theorem only under its matching assumptions [Mildner et al., 2025b], not from the size of the gap in this figure. A larger, higher-capacity model is needed to exhibit the effect at the scale reported by the source paper (sec. 8.19).

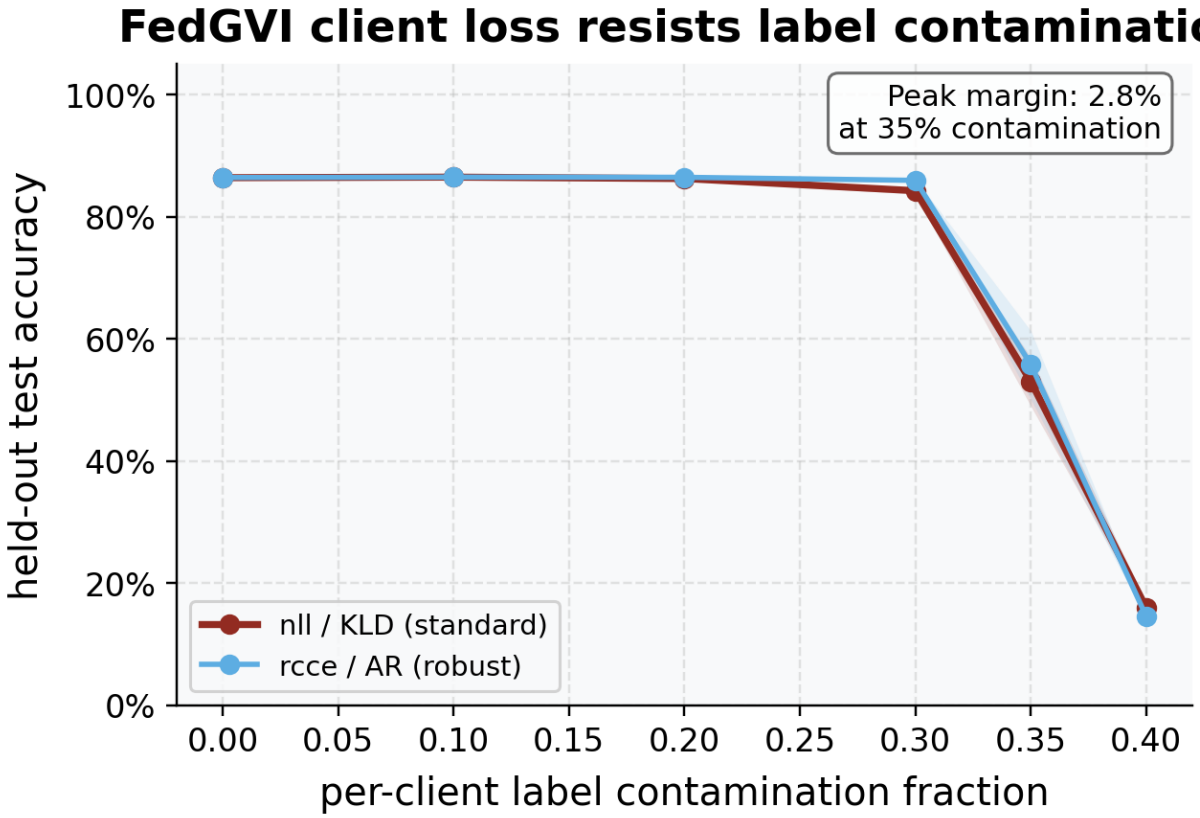


Figure 16: Held-out classification accuracy of the federated Bayesian baseline. Source relation: original project FedGVI complement; estimand: clean held-out accuracy fraction; uncertainty: seed-level bootstrap interval. The *logistic-regression* baseline (5 clients, 200 points per class per client, gradient-descent point-estimate weights — no posterior covariance is computed for this anchor) as a function of per-client label-contamination fraction. x-axis: contamination fraction (fraction of each client's labels flipped); y-axis: held-out classification accuracy on a clean test set, averaged over 64 independent seeds. The standard configuration (`nll` loss / `KLD` regularizer) and the robust FedGVI configuration (`rcce` loss / `AR` regularizer, $q = 1.00$) are shown as separate curves; shaded bands show seed-level 95% bootstrap intervals. The two curves are close at low-to-moderate contamination, separate over the moderate-to-high range (peak margin at 0.35 contamination), then reconverge at the highest swept level, where both decline sharply and neither curve reliably leads — that level is included rather than omitted, since it is the one part of the sweep that does not favor the robust client. Note: this figure plots the NumPy logistic-regression anchor, **not** the separate PyTorch deterministic MLP of the final paragraph (whose 16-hidden-unit, $\beta = 0.5$ configuration is an executed point-mass-family complement). The recovery identities establish compatibility at the named limit; the per-client bounded-influence result belongs to the cited FedGVI theorem under its matching assumptions, distinct from the server-side heuristic reweighting shown in the robustness results. Each point and interval is computed across 64 independent seeds.

fig. 16 is per-client empirical evidence. Its recovery identity and the source FedGVI theorem have separate roles; neither comes from the aggregation-level statistics of sec. 7.5.1. The three robustness axes — the source-conditional per-client update here, the complementary sharp server-side heuristic of sec. 7.5, and the conservative variational server rule of sec. 11 — remain distinct throughout. Only the per-client axis carries a source-conditional bounded-influence result; the variational server axis carries a raw effective-weight bound (sec. 11.3), not an estimator-level guarantee.

PyTorch deterministic-MLP complement (executed). As a generative-model-free complement, the analysis pipeline instantiates FedGVI in a deterministic point-estimate MLP — generalized variational inference with a point-mass variational family: Linear→ReLU→Linear→softmax with 16 hidden units, the density-power β-loss at $\beta = 0.5$, trained for 200 Adam steps per client across 5 clients — and fuses per-test-point softmax predictions with `robust_aggregate` at `robustness = 0.5` (`fedference.bnn_baseline_torch.run_bnn_torch_experiment`, run under PyTorch 2.12.1). Every number here is executed, not assumed: the consensus is a valid probability simplex (maximum deviation from unit mass 2.22e-16 over the test set)

and is bit-identical across repeated seeded runs (deterministic: Yes). Held-out consensus accuracy at contamination 0.40 is 0.558 for the $\beta \to 0$ standard client and 0.545 for the $\beta = 0.5$ robust client — this is the same 0.40-contamination endpoint where the NumPy baseline above also loses its separation (a single seed here, versus the 64-seed mean above), so the small gap is consistent with, not in tension with, that figure's genuine mid-range margin: both axes show the same qualitative collapse-together behavior at the sweep's most extreme point. This run confirms that the server-side aggregation API transfers to this neural-network setting and produces a valid, deterministic consensus; it does not establish model-class universality or that the client-side β-loss's robustness margin transfers at this scale; the certified NumPy logistic-regression baseline above remains the axis's rigorous evidence. When PyTorch is not installed the pipeline records a skipped status with unavailable-value sentinels; a complete certified build therefore installs the `torch` optional extra (sec. 10).

8 Discussion: what the evidence supports

The 9 studies probe distinct parts of one categorical, factor-based framework rather than repeating a single benchmark. Their common result is narrow but useful: the client construction contains its standard-Bayes limit and the server has a named project log-linear-pool recovery corner, and behavior away from those limits can be measured under declared contamination, sampling, and model assumptions. The recovery identities are formal; the performance results are conditional simulation evidence.

8.1 The recovery limit is the formal anchor

The coherence of the framework rests on the KL/NLL client limits and the zero-robustness project-pool identity (eq. 7, eq. 6). This is an identity of the stated categorical implementation, not an asymptotic claim about every generalized Bayesian model or a reconstruction of the complete source protocol. The server aggregator returns the log-linear pool (eq. 6) with maximum deviation 0; the generalized posterior returns the closed-form Bayes update with deviation 5.55e-17; and the Rényi divergence, β-loss, and robust cross-entropy recover their KL/NLL limits with residuals 0, 0, and 0. Theorem 5, Corollary 6, Lemma 3, and Proposition 4 state the formal limits; the recovery checks in sec. 7.1 are the executable falsification harness.

8.2 What the study suite jointly shows

Read together, the suite separates into two kinds of result: identities that hold exactly at the standard corner, and performance contrasts that are explicitly conditional on the operating regime. The distinction is the point of the joint reading — it says which claims travel and which are tied to the declared world.

Studies 1–3 sit at the standard corner. In the reduced categorical source-mechanism analogue protocol, belief sharing lowers mean variational free energy by $\Delta\bar{F} = 3.3109$ nats (sec. 7.2, fig. 8); Dirichlet language learning reduces KL from 3.4231 to 0.0027 nats (sec. 7.3, fig. 10); and Bayesian model reduction accepts the redundant-pruning candidate while rejecting the supported-column control (sec. 7.4, fig. 11). These are mechanistic checks, not evidence that the implementation matches every detail of the source simulations.

Study 4 steps away from the corner by holding the hidden state and attack target fixed while redrawing matched contaminated colonies. The standard pool degrades across the declared rate grid. The server-side heuristic is not uniformly better: its robust members are similar to or slightly below the standard pool at low contamination, then the selected pooled display member separates in its favor under severe contamination. At the largest swept rate, AR reaches 0.9880 against 0.6928 for the standard pool (tbl. 3; the sweep figure fig. 12 plots the separate matched-trial gallery colony, whose largest-rate separation is smaller). This regime dependence is a result, not a nuisance to be hidden: robustness can cost efficiency when the attack is weak and pay off when the declared contamination is severe.

The structural extension studies (Studies 5–9) sharpen the conditional half of that split rather than adding further corner checks, and they are reported with their negative contrasts intact. Communication is not uniformly beneficial: the cross-study summary (fig. 27) reports each study's federation benefit in its native units with seed-level intervals — several studies clearly positive, the moving-world EFE and two-level hierarchical studies approximately zero — so pooling helps when views are complementary and can be unnecessary or mildly costly when the agents already agree. Adding hierarchical depth is held to the same standard — the two-level stack does not beat the flat baseline on location accuracy (the paired gap is a small, statistically reliable cost) (sec. 13.9), earning its place only by additionally resolving the context latent above chance. The joint lesson is therefore not that federation or depth is always worth its cost, but that the suite measures the regimes in which each one is.

8.3 What this simulation identifies—and what it does not

The primary robustness estimand is the matched-trial mean difference in consensus accuracy conditional on the seeded true state and attack geometry. The 960 paired trials quantify Monte Carlo variation for that estimand; they do not average over hidden states, attack targets, adaptive adversaries, or real deployments. The cross-study layer reduces its matched trials within seed before seed-level summaries, so clients and within-seed trials are not silently counted as independent replicates. This follows simulation study guidance to declare the estimand and Monte Carlo unit explicitly [Morris et al., 2019, Koehler et al., 2009], and the bootstrap interval follows the declared resampling unit rather than treating nested observations as a flat sample [Loy and Korobova, 2021].

The consequence is a more informative claim boundary. The sweep supports a conditional statement about this contamination mechanism and these categorical beliefs; it does not establish universal Byzantine tolerance, calibration, or truth recovery. The result also does not identify a single universally best robustness parameter independent of the operating regime: the standard pool is preferable at the low-contamination cells in this run, while at least one robust member has a BH-rejected positive contrast in the declared high-contamination verdict; the tied display set and deterministic tie-break are reported separately. Those are precisely the conditions a future adaptive or deployment study must vary.

8.4 The robustness verdict is conditional and statistically qualified

The headline comparison is computed by the statistics module (sec. 7.5.1), not typed into the prose. Across 960 matched trial replicates nested within the fixed seeded world at contamination rate 0.800, each robust server member is compared with the standard pool by a Wilcoxon signed-rank test [Wilcoxon, 1945] and the declared family of p-values is adjusted by Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995]. A method wins only when the adjusted null is rejected and its effect is positive, here at $q = 1.11 \times 10^{-158}$ with effect size 1.0000 (large). The observed-effect power and prospective sample-size calculation are planning quantities, not independent confirmation of the result.

The paired design is appropriate for the controlled comparison because each condition shares the seed, true state, and attack geometry. Its interpretation remains limited: the signed-rank test concerns the distribution of paired differences, not equality of raw means, and BH controls expected false-discovery proportion within the declared family rather than the probability of any false positive. The confidence intervals are percentile bootstrap intervals over the matched-trial unit. Seed-level review-grid results are reported separately and do not treat the primary trial count as a population of independent worlds. These qualifications make the verdict narrower, but also make it reproducible and falsifiable.

8.5 Three robustness axes remain separate

The unifying narrative would be dishonest if it let the aggregation heuristic inherit the guarantees of the client update or variational server objective. The **client-side** β/rcce generalized-Bayes update is the FedGVI-faithful axis [Mildner et al., 2025b]: it is derived from the generalized-Bayes objective (eq. 1), limits to NLL/Bayes as its loss parameter tends to zero, and carries the loss-specific bounded-influence result (eq. 4, eq. 5). The **server-side** `robust_aggregate` divergence-reweighting is a complementary heuristic whose proven property is the recovery limit only (eq. 7). The **variational server** axis is objective-backed and supplies a proven raw effective-weight bound, with conservative accuracy behavior. No figure, statistic, or sentence transfers a client guarantee to the heuristic or a variational weight bound to the heuristic's accuracy verdict; sec. 8.13 states the boundary in full.

8.6 Accuracy and effective-weight control can be traded explicitly

The F_λ family (sec. 11.5) supplies empirical grid evidence that accuracy and effective-weight control need not be a fixed binary choice within the tested objective family. At $\lambda = 1.0$ the implementation recovers the current variational objective bit-for-bit; lower explored temperatures sharpen the consensus toward the heuristic while preserving the stated raw-weight bound under its assumptions. This is not a derivation of the heuristic from an objective. The open problem is to identify an objective whose minimizer is both competitive across contamination regimes and accompanied by a theorem that survives beyond the present categorical construction.

8.7 Why the boundary matters downstream

The practical lesson is not that every robust rule should replace belief sharing. It is that a multi-agent active-inference system can expose separate controls for client updating, server reweighting, and objective-backed consensus, while retaining a tested route back to the ordinary pool. That separation tells a builder what can be promised: exact recovery at the named corner, conditional empirical behavior under the declared attack, and no silent transfer of a client-side or variational theorem to a different server heuristic. The result is a usable research contract for extending belief-sharing systems without confusing a simulation win with a general robustness guarantee.

8.8 Related work: active inference, federated Bayes, and the scoped bridge

This work sits at the boundary between two research communities with different objects of inference. Each cited thread contributes a real component; this paper extends the intersection without claiming that either literature is exhausted. The positioning is therefore against the reviewed sources and their assumptions, not against an absolute claim about everything the field has or has not done.

8.9 Pre-modern probability, inverse probability, and collective judgment

The pre-modern sources in this manuscript are not evidence that early probability theorists anticipated KL minimization or federated learning. They serve a narrower role: they show that the paper's recurring problems — expected uncertain outcomes, inverse inference from effects to causes, utility-sensitive action, and collective judgment — are old problems now implemented with modern variational machinery.

The probability-calculus line begins with Pascal and Fermat's correspondence on the problem of points [Pascal and Fermat, 1654], Huygens' printed treatment of reasoning in games of chance [Huygens, 1657], Montmort's combinatorial analysis of games [Montmort, 1708], Bernoulli's *Ars conjectandi* and its law-of-large-numbers logic [Bernoulli, 1713], and de Moivre's systematic probability textbook [de Moivre, 1718]. For this paper, their relevance is not that they contain active inference, but that they make expectation and uncertain evidence objects of calculation.

The inverse-probability line is closer to the manuscript's formal spine. Bayes and Price frame the problem of inferring an unknown chance from observed events [Bayes, 1763], and Laplace generalizes the probability of causes given events [Laplace, 1774]. Read beside modern active inference, these sources support the vocabulary of generative-model inversion: data are effects, latent states or parameters are causes, and a posterior belief reconciles prior commitments with observed evidence. That is a conceptual bridge only; the actual identity with the FedGVI recovery corner is the modern result of sec. 5.8 and sec. 6.

Decision and aggregation enter through Daniel Bernoulli's expected-utility treatment of risk [Bernoulli, 1738] and the voting-theoretic work of Borda and Condorcet [Borda, 1784, Condorcet, 1785]. They help name the problem this paper revisits in a categorical active-inference colony: local judgments must be combined, and the chosen aggregation rule encodes assumptions about competence, independence, weights, and stakes. Those assumptions are exactly what the modern log-pool and robustness literature make explicit.

8.10 Active inference: generative agents, EFE, and colonies

The free-energy principle [Friston, 2010] and its discrete state-space process theory [Friston et al., 2017] gave the field a unified account of perception and action: hand-built generative models with $A/B/C/D$ matrices, variational free energy for inference, and expected free energy for principled action selection. The discrete state-space synthesis [Da Costa et al., 2020] and toolboxes turned that account into scalable machinery — the pymdp discrete-state library [Heins et al., 2022] and the RxInfer reactive message-passing engine [Bagaev et al., 2023]. *Yes:* this substrate is exactly what we reimplement, and our EFE decomposition into risk and ambiguity (eq. 18, eq. 19, fig. 7) is the community's own action-selection objective. *And:* we add a robustness knob to its belief-fusion step without leaving the framework.

The fusion operator itself is not new as a mathematical object. Under the explicit categorical posterior-log-potential and fixed-weight bridge of sec. 5.8, the message-combination term of Friston's belief-sharing equation is represented by a logarithmic opinion pool [Genest and Zidek, 1986, Genest et al., 1986] and a product-of-experts consensus [Hinton, 2002]. That limited representation is not a statement that the complete source protocol is a log pool. Abbas' KL view of linear and log-linear pools makes the same point in scoring-rule language: log-pooling can be justified as a KL aggregation rule for expert distributions, not as a contamination-robust estimator [Abbas, 2009]. The Bayesian Committee Machine adds the distributed-learning analogue: independent estimators trained on data subsets can be combined by a product-style Bayesian rule, with assumptions about conditional independence and prior accounting made explicit [Tresp, 2000]. Fully Bayesian aggregation gives the social-choice counterpart: geometric pooling of beliefs is singled out by dynamic Bayesian rationality conditions, not by robustness to contaminated reports [Dietrich, 2021]. That classical literature is useful precisely because it names the hidden commitments — shared support, weight choice, prior accounting, and external-Bayes coherence — that are easy to overlook when the same operation appears as a colony update.

A parallel thread studies how active-inference agents coordinate as ensembles — collective behavior and surprise minimization across a colony [Heins et al., 2024], epistemic communities [Albarracin et al., 2022], and collective intelligence [Kaufmann et al., 2021]. The belief-sharing thread [Friston et al., 2024] is one member of this family: agents reach consensus by exact-Bayes fusion of one another's beliefs, and the colony belief-sharing scenario we implement as a reduced categorical standard-Bayes-limit analogue (eq. 10, sec. 7.2) is related to its worked example. Structure learning rounds out the picture: active inference with Bayesian optimal design selects and reduces models within the same frame [Smith et al., 2022], and post-hoc Bayesian model reduction [Friston and Penny, 2011] is the engine behind our emergence study (eq. 16, sec. 7.4). The cited active-inference line has rich generative models, principled action, and ensembles that coordinate by sharing beliefs. In the sources reviewed here, the belief-fusion rule is not systematically characterized under explicit contamination or intentionally wrong broadcasts, nor connected to the robustness theory used here. Fusion is treated as exact-Bayes and trusting in that scoped comparison.

Friston et al. [2024] crystallized these ideas into three worked simulations: (1) communicating-colony free-energy convergence, (2) Dirichlet language acquisition, and (3) Bayesian model reduction structure emergence. We use reduced categorical standard-Bayes-limit analogues of all three mechanisms (sec. 7.1) under the declared protocol before building the robust extension. This is a mechanism-level comparison, not an exact source-protocol or figure reproduction.

8.11 Robust and federated Bayes outside active inference

Federated learning aggregates models trained on decentralized data [McMahan et al., 2017], and the probabilistic-federation line recasts that aggregation as variational inference — partitioned variational inference [Ashman et al., 2022] and its federated predecessor [Bui et al., 2018]. This is a different use of the word *federated* from Friston et al.'s federated inference: Friston federates hidden-state beliefs among agents sharing a world model, while machine-learning federated learning usually federates parameter or predictive-model updates over decentralized datasets. The bridge claimed here is therefore algebraic and variational — a shared normalized product/log-pool operator at the KL/NLL recovery corner — not a claim that either source paper already solved the other's problem.

Robustness, meanwhile, has a mature Bayesian theory: general Bayesian updating through a loss [Bissiri et al., 2016], Gibbs posterior inference [Jiang and Tanner, 2008], safe learning-rate selection under misspecification [Grünwald, 2012], coarsened posteriors for robustness to exact-data conditioning [Miller and Dunson, 2018], divergence-criteria posterior updating [Jewson et al., 2018], Bayesian misspecification asymptotics [Kleijn and van der Vaart, 2012], the optimization-centric generalized variational inference view [Knoblauch et al., 2022], recent closed-form characterizations of unrestricted generalized variational objectives [Nguyen and Westerhout, 2026], and the bounded-influence losses that make updates robust — density-power and gamma-divergence estimation [Basu et al., 1998, Fujisawa and Eguchi, 2008, Ghosh and Basu, 2015], robust-divergence variational inference [Futami et al., 2018], and generalized cross-entropy [Zhang and Sabuncu, 2018]. Huber and Ronchetti supply the robust-statistics vocabulary of influence and breakdown [Huber and Ronchetti, 2009], which we use as vocabulary for boundedness and empirical failure modes rather than as a theorem transfer. FedGVI [Mildner et al., 2025b] is the synthesis we use per agent: a robust generalized- Bayes objective with client and server divergence choices. This community provides decentralized aggregation and theorem-backed results in its stated settings. The cited papers do not evaluate active-inference POMDP belief consensus — the generative-model-bearing, action-selecting setting where beliefs drive behavior.

The 2025 preprint on convergence rates under prior misspecification [Mildner et al., 2025a] sharpens the current GVI context: bounded divergences can support concentration and rates under explicit assumptions, but the result does not transfer automatically to this repository's finite categorical state space or to its server-side aggregation heuristic.

The federated-learning robustness literature also supplies important negative space for this paper. Byzantine-tolerant gradient aggregation [Blanchard et al., 2017], geometric-median robust aggregation [Pillutla et al., 2022], and divergence-weighted gamma-mean aggregation [Li et al., 2022] attack corrupted client updates directly. A recent Bayesian robust-aggregation preprint likewise models unknown client honesty for federated model updates [Karakulev et al., 2025]. Those are close comparators for adversarial federation, but the object being aggregated differs: they aggregate model-update vectors or posterior measures, while this paper attacks categorical belief fusion and generalized-Bayes client updates. Robust subset-posterior combination is another nearby Bayesian route [Minsker et al., 2017], but it combines posterior measures across data shards rather than active-inference belief broadcasts with a shared latent state. We use those sources to position the problem, not to import their guarantees into `robust_aggregate`.

8.12 The specific bridge added here

The gap this manuscript fills is robust, generalized-Bayes belief fusion for active-inference ensembles. Concretely:

- **Per-agent bounded-influence updates.** FedGVI-faithful β/rcce updates that carry provable bounded influence (eq. 4, eq. 5).
- **Server-side divergence-reweighting heuristic.** A complementary aggregation heuristic (eq. 7) with a stated recovery limit, positioned alongside robust federated aggregation work without borrowing its theorems.
- **Provable recovery of the standard-Bayes client limit.** The client KL/NLL loss limits recover Bayes, and the separate zero-robustness server identity returns the project log-linear pool (eq. 6, eq. 3, eq. 7). Under the stated categorical bridge, that pool specializes only the Eq. 7 message-combination term [Friston et al., 2024].
- **Additional executed studies beyond the Friston et al. (2024) baseline.** The three Friston simulations motivate source-mechanism analogues rather than serving as source-protocol recovery checks; beyond them we contribute: (4) a contamination sweep in which a fraction of colony members are adversarial or misspecified (sec. 7.5); (5) a moving-world scenario with active expected-free-energy movement (sec. 13.7); (6) a two-level hierarchical POMDP (sec. 13.9); (7) a three-level hierarchical POMDP (sec. 13.11); (8) a sensitivity sweep over acuity × colony size (sec. 13.13); and (9) finite-grid parameter recovery for the declared acuity family (sec. 14).
- **Rigorous statistics.** Every verdict is produced by matched-pairs Wilcoxon signed-rank tests [Wilcoxon, 1945, Fay and Proschan, 2010] deflated with Benjamini–Hochberg FDR [Benjamini and Hochberg, 1995], reported with bootstrap confidence intervals [Efron and Tibshirani, 1993], rank/effect-size caveats [Nakagawa and Cuthill, 2007], and an observed-effect design-power approximation — none of which appear in Friston et al. (2024).
- **An objective-backed server-side aggregator with redescending weights.** The `variational_aggregate` rule descends a stated aggregation free energy (eq. 8) monotonically; any converged fixed point is coordinatewise stationary. The rule carries a proven raw effective-weight bound and empirical redescending response (fig. 15), while leaving open whether the sharper reverse-KL server heuristic is itself the closed-form minimizer of an equally defensible objective.

The contribution is the bridge: an explicit connection between active-inference belief consensus and robust federated generalized Bayes, with the standard pool recovered exactly at the corner and the additional studies (the contamination sweep plus several structural extensions) showing how the bridge behaves beyond the Friston et al. baseline.

8.13 Limitations and claim boundaries

The boundaries here define the contribution. The central goal is to show — concretely and testably — that the categorical generalized-variational construction inspired by FedGVI [Mildner et al., 2025b] has standard-Bayes client limits and a project-local log-linear-pool server identity. Under the explicit bridge in sec. 5.8, the latter specializes Friston et al.'s Eq. 7 message-combination term [Friston et al., 2024], not the complete source protocol; behavior away from those limits is evaluated only under declared simulation conditions. The exact identities are formal; performance and deployment claims remain conditional. Items beyond that boundary are named as future work.

8.14 Three robustness axes: theorem, heuristic, and objective

Robustness enters in three places with unequal standing — a distinction surfaced by an adversarial review of this work and preserved deliberately throughout (sec. 3.3).

1. **Client-side, source-theorem-backed.** A robust per-agent generalized-Bayes update with a bounded loss — β-loss [Basu et al., 1998] or generalized cross-entropy [Zhang and Sabuncu, 2018] — inside the generalized posterior. Density-power and gamma-divergence Bayes make clear that this is a loss/divergence-specific robustness property, not an automatic property of every generalized posterior [Fujisawa and Eguchi, 2008, Ghosh and Basu, 2015]. It is derived from the generalized-Bayes objective [Bissiri et al., 2016, Jewson et al., 2018, Knoblauch et al., 2022], provably limits to NLL/Bayes (hence to the standard pool) as the loss parameter goes to zero (eq. 4, eq. 5), and is the theorem-bearing axis under the matching loss, divergence, and regularity assumptions of FedGVI. The manuscript does not re-prove that theorem for every possible categorical data-generating process. **This is the client-side mechanism implemented and evaluated here.**
2. **Server-side, heuristic.** The divergence-reweighting aggregator `robust_aggregate`. Its positive formal property is the recovery limit — robustness 0 equals the standard log-linear pool (eq. 7). A scoped no-go rejects one declared separable objective class; it is *not* a closed-form minimizer of a FedGVI objective, and we never claim it inherits the bounded-influence bound. An empirical characterization (sec. 14.2) makes this boundary concrete: the heuristic has a *finite* measured breakdown point (a colluding majority captures it), a witness against unconditional server resistance rather than a transferred client theorem. The robustness sweep reports it as a complementary heuristic, and fig. 13 is labeled accordingly. Robust federated optimizers such as Krum and gamma-mean aggregation motivate the adversarial client problem [Blanchard et al., 2017, Li et al., 2022], but they do not certify this belief-pooling heuristic. It has BH-rejected positive contrasts in the configured accuracy verdict in sec. 7.5.1, while the declared mechanism gallery and conditional-world study retain reversals that prohibit a universal accuracy claim.
3. **Server-side, objective-backed (conservative).** The variational aggregator `variational_aggregate` of sec. 5.8.2, derived in sec. 11. It runs exact closed-form block updates that descend the stated free energy eq. 8 monotonically on each block, shares the recovery corner (eq. 7) with axis 2, and — unlike axis 2 — carries a proven raw effective-weight bound with empirical redescending behavior (fig. 15). Its cost in the declared comparison is conservatism: the entropy regularization yields a more diffuse consensus, and the method does *not* win that configured peak-accuracy verdict. It closes the "is there *any* objective-backed server rule with raw-weight control" gap; it does not retroactively endow the sharp axis-2 heuristic with an objective.

The honest state is therefore a triangle, not a binary: axis 1 is source-theorem-backed under matching assumptions, axis 2 has conditional wins and reversals *without* a server-side objective, and axis 3 is objective-backed *but* conservative. No experiment attributes axis 2's accuracy win to a theoretical guarantee, and none attributes axis 3's weight bound to a peak-accuracy claim. The remaining open problem — an objective whose minimizer is the *sharp* reweighting itself, which would combine accuracy and guarantee in one server rule — is named future work in sec. 8.17; the extended methods that broaden the toolkit without touching these claims are cataloged in sec. 12.

8.15 Scope boundaries that the evidence does not cross

- **The bridge is a recasting, not an upstream claim.** Friston et al. [Friston et al., 2024] supply a belief-sharing operator for agents with a shared world model; Mildner et al. [Mildner et al., 2025b] supply robust federated generalized variational inference. Neither paper claims the other. The contribution here is to recast the former as the tested KL/NLL zero-robustness recovery corner of the latter-inspired construction, then measure what changes when robust losses and server reweighting are introduced.
- **Historical sources are conceptual, not formal support.** The pre-modern sources added in sec. 8.9 support a genealogy of expectation, inverse inference, utility, and collective judgment [Huygens, 1657, Bayes, 1763, Laplace, 1774, Bernoulli, 1738, Condorcet, 1785]. They do not support any claim about KL, NLL, product-of-experts training, FedGVI, or `robust_aggregate`. Those claims rest only on the modern cited formalism and the tests reported in sec. 7.1.
- **Single-machine federation.** Federation is validated with queue transport, a single-machine OS-process helper, and loopback TCP: agents serialize beliefs, the server aggregates, and consensus is broadcast back without changing the mathematics. The socket path now exercises frame integrity and file-backed digest-verified replay validation; an optional SQLite round-ID guard also survives local process restarts. These controls are still not a substitute for cross-host deployment, identity-bound mTLS, a shared multi-host replay domain, discovery, long-running worker orchestration, or fault tolerance; those steps are scoped as future work (sec. 8.17).
- **GPU / PyTorch Bayesian-neural-network experiments at the original FedGVI scale.** The bounded-influence result is anchored here by a small NumPy logistic-regression baseline (fig. 16), not the RTX-class runs of the source paper [Mildner et al., 2025b]. Even on that anchor the robust client's margin over the standard client is non-monotone rather than uniform: it opens in the moderate-to-high contamination range and then vanishes at the most extreme swept level, where both configurations collapse together with no reliable ordering (sec. 7.6). That terminal convergence is reported, not trimmed; the axis's rigorous standing rests on the recovery identities and the FedGVI theorem under its matching assumptions, not on the size or monotonicity of the empirical gap in this figure.
- **Classification baselines are point estimates, not full posteriors.** The NumPy logistic-regression baseline and the PyTorch MLP complement both use point-estimate weights (no posterior covariance is computed). A genuine mean-field variational family over the weights — diagonal-Gaussian $q(w)$ with a closed-form KL and a Monte-Carlo ELBO — is implemented as a tested primitive (`bnn_variational_torch.VariationalMLP`, recovering the point-estimate net exactly as its posterior variance vanishes), but the full paper-faithful FedGVI classification lane (that variational family trained under the contamination sweep, at GPU scale) is not: MCMC, non-diagonal structure, and stochastic-weight averaging remain unimplemented, and the full-posterior regime remains unverified.
- **Hierarchical depth does not by itself improve the base task.** The two- and three-level stacks (sec. 13.9, sec. 13.11) run alternating L1/L2(/L3) inference end-to-end and resolve their added context latents above chance, but on the shared location task they do not beat the flat baseline — the paired location-accuracy gap is a small, statistically reliable negative gap at the reported seed count (fig. 25). Depth is therefore validated as executable and consensus-preserving, not as an accuracy improvement; whether a richer policy-and-horizon task family rewards hierarchy is future work (sec. 8.20).
- **Discrete POMDP only — continuous or hybrid state spaces are not addressed.** This work is discrete-categorical only, matching the community's worked POMDP example [Da Costa et al., 2020, Friston et al., 2024]. The limit-as-proof contract is validated off the categorical case only for a one-dimensional Gaussian-mean conjugate slice; continuous-state active inference and Gaussian belief-sharing colonies remain untested.
- **Temperature and divergence calibration are fixed, not learned.** Coarsened posterior and safe-Bayes theory make clear that the learning-rate/temperature is part of the inference rule under misspecification [Miller and Dunson, 2018, Grünwald, 2012, Kleijn and van der Vaart, 2012]. This manuscript validates the configured losses and divergences, but it does not learn an optimal coarsening radius, temperature, or divergence schedule across agents.
- **Real multi-machine federation, networking, and privacy cryptography.** Federation here is *mathematical* (factor aggregation), not infrastructural — unlike communication-efficient or Byzantine-robust federated learning [McMahan et al., 2017, Blanchard et al., 2017].
- **New linguistic theory.** The language-acquisition study (sec. 7.3) reproduces the Dirichlet count mechanism (eq. 14) mechanically; it does not extend the linguistics.

8.16 What the statistics can and cannot claim

The verdict is a paired comparison at a single high contamination rate (sec. 7.5.1, tbl. 5). This concentrates power where the effect is largest, which is honest about *where* robustness pays off but does not characterize the full contamination curve as a continuous function; the per-rate table (tbl. 4) reports the rest of the sweep without claiming family-wide significance beyond what BH-FDR [Benjamini and Hochberg, 1995] certifies. BH-FDR controls expected false discovery proportion within the declared family, not the chance of any false positive. The matched-pairs Wilcoxon tests are rank tests under paired-difference assumptions [Wilcoxon, 1945, Fay and Proschan, 2010], not assumption-free proofs about raw means. Confidence intervals are percentile bootstrap [Efron and Tibshirani, 1993], not analytic, and inherit that method's small-sample caveats at the lowest trial counts. The power values are observed-effect design approximations useful for confirmatory sample-size planning; they are not independent evidence for the verdict and do not cover model-specification uncertainty outside the seeded simulation harness [Wasserstein and Lazar, 2016].

8.17 Future work: testing the open boundaries

The staged research program follows directly from the boundaries in sec. 8.13. It separates public-library reproducibility, server theory, portable source-protocol work, task generalization, and deployment validation so that success in one lane cannot silently certify another.

The order is deliberate. Simulation-study guidance recommends declaring the estimand, data-generating mechanism, and Monte Carlo precision before treating replication as evidence [Morris et al., 2019], while Monte Carlo error should be reported separately from an interval or a hypothesis test [Koehler et al., 2009]. For nested agents, trials, and seeds, the resampling unit must respect the dependence structure [Loy and Korobova, 2021]. Accordingly, the phase plan records a primary unit and a falsifier for each extension; a larger sample or more elaborate diagram is not itself a stronger claim.

The implementation registry also separates smoke, pilot, and confirmatory profiles. Pilot worlds select budgets and calibration settings but never enter confirmatory intervals or headline values. Each completed run must bind its source bundle, configuration, dataset bytes, device, checkpoints, outputs, and completion status into a verifiable receipt. A negative or null scientific result remains a valid citable outcome when those implementation and provenance gates pass; it blocks the intended positive claim, not the software release.

A separate Friston protocol lane will resolve a machine-readable parity matrix before reconstructing source experiments in Python. Until every required source parameter, routine, unit, and estimand is recovered, that lane will be described as a paper-constrained reconstruction rather than an exact replication.

8.18 Make the sharp server heuristic variational

The asymmetry between the robustness axes (sec. 3.3) is the most consequential open problem. The client-side β/rcce update carries a derived, loss-specific bounded-influence result under the matching assumptions; the sharp server-side `robust_aggregate` carries only its recovery limit (eq. 7); and the new variational aggregator (sec. 5.8.2, sec. 11) supplies a *rigorous* server-side rule — exact block updates descending the stated free energy eq. 8 through non-increasing block updates, with a proven raw effective-weight bound and the same recovery corner. What it costs is conservatism: it is the maximum-entropy consensus, not the sharp accuracy-maximizer. The remaining open problem is therefore sharper than before — to write down a generalized variational objective in the FedGVI family [Mildner et al., 2025b], informed by recent closed-form GVI characterizations [Nguyen and Westerhout, 2026], logarithmic-pool weighting theory [Carvalho et al., 2023], and robust divergence-weighted federated aggregation [Li et al., 2022], whose closed-form minimizer is competitive with the empirical reweighting across declared contamination regimes. That would combine axis 3's effective-weight bound with axis 2's empirical sharpness in one server rule. The recovery corner and the variational objective together supply two boundary conditions any such unification must satisfy.

Any empirical choice of `robustness` or `entropy_weight` will be made on separate calibration worlds using a proper log score, then frozen before confirmatory evaluation. This guards against selecting an apparent leader with evaluation truth and preserves null or reversed confirmatory outcomes. The comparison family will include logarithmic and linear pools, the current heuristic, the variational family, and a centered-log-ratio geometric-median control; none inherits a parameter-space robust-federated-learning guarantee merely by operating on beliefs.

8.19 Promote the baseline to original FedGVI scale

The bounded-influence result is anchored locally by the small NumPy logistic-regression baseline (fig. 16). Promoting it to the GPU-scale Bayesian-neural-network experiments of the source paper [Mildner et al., 2025b] — the experiments deferred here (sec. 8.13) — would test whether the per-client robustness curve holds at the model capacity and contamination regimes where federated learning [McMahan et al., 2017] actually operates, and would connect the discrete-POMDP result to the partitioned-VI line [Ashman et al., 2022, Bui et al., 2018]. The planned comparison would also require posterior-parameterization parity with Bayesian neural-network work, rather than treating the current deterministic point-mass MLP as a posterior [Mildner et al., 2025b].

The portable lane preserves the source protocol's site factors, client cavity, and factor-replacement update in natural coordinates. It distinguishes a locally budgeted CPU/MPS profile from an exact source-scale CUDA profile that remains external until suitable hardware is available. FashionMNIST anchors protocol parity, while MNIST and KMNIST test portability. A separate source-bound tabular pack will report proper-score effects per licensed dataset, with training-only preprocessing and byte-, split-, and license-level provenance; nested seeds will not be treated as independent datasets.

8.20 Extend hierarchical federation beyond the current stack

The governing caveat here is that added depth must be shown to *earn* generalization, not merely to execute: on the current sentinel task the deeper stacks match rather than beat the flat baseline on location (sec. 8.15). The extension therefore has two distinct fronts — carrying the recovery contract to deeper stacks, and finding a task family in which depth actually pays.

The 2-level hierarchical POMDP (sec. 13.9) couples location inference to a single global context. The generic N-level architecture (:func:`fedference.pomdp.build_nlevel_world`) has already been exercised with a 3-level stack (sec. 13.11, sec. 13.12): a meta-context variable (L3) gates the context prior (L2) which in turn gates the location prior (L1). The empirical-prior top-down messages (eq. 32, eq. 33) remain valid variational steps at every depth, and the log-linear-pool federation at each level is bit-identical to the in-process result (Proposition 12). The natural next question is whether the limit-as-proof contract of sec. 8.13 survives still deeper hierarchies: does the recovery corner (context prior $\to$ uniform) remain checkable to machine precision when the L2 $\to$ L1 message is itself a function of an L3 belief coupled to an L4 belief, and can message-passing engines such as RxInfer [Bagaev et al., 2023] carry the generic alternating-minimization at scale? Structure learning already answers the dual question — how *deep* the model should be — for the top level: hierarchical Bayesian model reduction (sec. 14.1) prunes a non-gating meta-context and keeps an informative one, so the depth is decided by the evidence rather than assumed. Extending that per-level reduction to a full breadth-and-depth search over the generic N-level stack is the natural continuation.

The next task family is deliberately controlled rather than merely deeper: partially observable Four Rooms and Key-Door will compare flat, oracle, learned, shuffled, and non-gating hierarchies at matched horizons and compute. Task is the higher-level replication unit. A gain in only one task will remain task-specific, and the larger campaign will not begin until the hybrid representation recovery gates pass.

8.21 Move from process transport to true multi-machine federation

Promoting federation from the current queue-backed, single-machine process and loopback-socket helpers to cross-host workers would retire the remaining deployment caveat of sec. 8.13 while preserving the bit-identical consensus property proved in Proposition 12. The `federation/` package and federation tests already establish the API contract: a server collects serialized beliefs from 5 worker channels, fuses them with `robust_aggregate` at robustness $c = 1.5$, and broadcasts the consensus back over response channels, with bit-identity verified at True. The loopback socket path adds optional pre-shared-key frame integrity and file-backed digest-verified replay validation. The next systems step is an explicitly local Docker

multi-node emulator with mTLS by default, HMAC compatibility, checkpoint/restart, and reproducible message-fault controls. It is not physical multi-host evidence. That later claim requires receipts from distinct hosts, deployment-grade key management, timeout policy, and long-running orchestration across process restarts, but it does not require changing the mathematics. Secure aggregation, differential privacy, and Byzantine tolerance are separate future threat models; transport integrity alone would not establish any of them [Blanchard et al., 2017, Pillutla et al., 2022].

8.22 Move beyond categorical state spaces

This work is discrete-categorical, matching the community's worked POMDP [Da Costa et al., 2020]. A first step is already in place: the closed-form Gaussian KL and Rényi divergences of sec. 12 show the divergence family — and its $\alpha \to 1$ recovery — carries over verbatim to Gaussian beliefs, scoped out of the categorical experiments. Continuous-state Gaussian generative models would test whether the limit-as-proof contract survives the move off categorical state spaces — whether the recovery corner remains checkable to machine precision when the belief simplex is replaced by a Gaussian belief, and whether message-passing engines such as RxInfer [Bagaev et al., 2023] can carry the robust update at scale. Extending the structure-learning study [Smith et al., 2022, Friston and Penny, 2011] into continuous models would, in parallel, test whether robust belief fusion and robust *structure* fusion compose.

A minimal executable fixture now gates a discrete dynamics context over continuous position and velocity, Gaussian observations, and bounded actions. It is a representation and recovery surface, not confirmatory task evidence. The full study must add discrete-only, continuous-only, and oracle-context controls, singular-covariance and outlier checks, and held-out posterior-predictive scoring before supporting a hybrid-task claim.

9 Conclusion: a recovery-tested bridge with bounded claims

Active Fedference makes a deliberately narrow bridge between two bodies of work that are usually discussed with different objects and different standards of evidence. Active inference describes agents that infer and act within generative models; federated generalized Bayes describes how losses, divergences, and variational objectives can make decentralized inference less sensitive to bad information [Friston et al., 2017, Da Costa et al., 2020, Bissiri et al., 2016, Knoblauch et al., 2022]. This paper does not dissolve those distinctions. It puts them into one categorical implementation, identifies the exact point at which they coincide, and then measures the consequences of moving away from that point.

9.1 The durable result is a recovery contract

The central contribution is the recovery-tested contract. The KL/NLL client limits of the declared generalized-Bayes construction recover the closed-form Bayes update, while the zero-robustness server branch recovers the project log-linear pool, with maximum measured deviations of 5.55e-17 and 0 (eq. 6; eq. 7). This is stronger than a verbal analogy because the limited identities are stated in the formalism, implemented in the core, and checked by executable invariants. Under the explicit shared-support, posterior-log-potential, and fixed-weight bridge, the server pool specializes Eq. 7's message-combination term; it is not an assertion that the active-inference and robust-Bayes literatures share all assumptions, objectives, deployment meanings, or the complete source protocol.

That distinction matters historically and methodologically. Logarithmic pooling has a substantial literature as an aggregation rule for expert distributions, including its connections to external Bayesianity, product-of-experts constructions, and KL-based opinion pooling [Genest and Zidek, 1986, Genest et al., 1986, Hinton, 2002, Abbas, 2009, Dietrich, 2021]. FedGVI contributes a generalized-Bayes perspective in which the loss and divergence determine what robustness means [Mildner et al., 2025b]. The contribution here is therefore a scoped recasting: the project's categorical log-linear-pool specialization is the non-robust corner of the implemented family under its stated bridge assumptions, and the corner becomes a testable boundary condition for future robust extensions. This does not identify the complete source belief-sharing protocol with FedGVI.

9.2 What the evidence establishes away from the corner

The standard-Bayes studies provide a necessary baseline rather than ornamental background. Communication changes mean free energy by 3.3109 nats, Dirichlet learning reduces KL from 3.4231 to 0.0027, and Bayesian model reduction selects the declared redundant-pruning candidate. These results recover the declared mechanisms that make belief sharing scientifically interesting while keeping the source relationship bounded to the stated categorical protocol [Friston et al., 2024]. The extension studies then show that communication is not automatically beneficial in every information geometry: disjoint views can make sharing valuable, whereas a complementary moving-world control can make additional pooling unnecessary or mildly costly.

The contamination study adds a second lesson. The server-side heuristic is regime-dependent: robust operating points can give up a little efficiency when contamination is weak and recover that cost when the declared attack is severe. At the most severe swept rate, the highest pooled robust mean reaches 0.9880 against the standard pool's 0.6928 (tbl. 3); at the verdict rate, the matched, BH-adjusted comparison of sec. 7.5 gives 0.9867 against 0.9021. The result is therefore an operating-point contrast, not a ranking that holds for every contamination rate, attack target, hidden state, or calibration regime. This interpretation follows the simulation-study principle that the data-generating mechanism, estimand, and Monte Carlo unit should be declared before a numerical result is treated as general evidence [Morris et al., 2019].

The three robustness axes give the result its proper meaning. The client-side generalized-Bayes update is the FedGVI-faithful, theorem-bearing axis under its matching loss and regularity assumptions. The server-side `robust_aggregate` is the sharp empirical heuristic; its proven property here is the recovery limit, not a transferred client-side bounded-influence theorem. The `variational_aggregate` is the conservative complement: it descends a stated aggregation objective and has a proven raw effective-weight bound, but its objective-backed control does not make it the peak-accuracy display leader. Separating these axes prevents a familiar failure in interdisciplinary work: a guarantee proved for one operator is silently attached to another operator because both are described with the same word, "robust."

9.3 Why the bridge matters for active inference

Belief sharing is not merely a communication convenience. In an active-inference system, a shared posterior can change expected free-energy calculations, model comparison, and subsequent action selection. A robustification at the fusion step can therefore alter behavior even when each local generative model is unchanged. Conversely, a server rule that suppresses an anomalous belief may also suppress a rare but correct observation. The relevant scientific question is not simply whether a robust curve rises; it is which assumptions about competence, independence, support, and action-relevant uncertainty the fusion rule encodes [Genest and Zidek, 1986, Tresp, 2000, Heins et al., 2024].

This is why the recovery corner is a useful organizing device. It gives a common reference behavior before robustness is introduced, makes the cost of a server intervention measurable, and lets later work compare new aggregation rules against an interpretable baseline. The bridge also keeps the distinction between inference and infrastructure visible. The current federation tests show that serialized beliefs can travel through the declared local transport and return a bit-identical consensus, but they do not turn a mathematical aggregation identity into a claim about secure, fault-tolerant, or privacy- preserving deployment [McMahan et al., 2017, Blanchard et al., 2017, Pillutla et al., 2022].

9.4 What remains unproved

The evidence does not establish universal Byzantine tolerance, truth recovery, calibration, or an optimal robustness parameter. The primary intervals are conditional on the fixed hidden state and attack target, and the nested trial and seed structure is reduced at the declared unit rather than treated as a larger independent sample [Koehler et al., 2009, Loy and Korobova, 2021]. The categorical state space is the object of the proof and experiment; continuous or hybrid state spaces are a separate mathematical extension. The neural classification complement uses point-estimate weights, so it does not establish full posterior FedGVI behavior at the scale of the source experiments.

These are not defects to be hidden by a more expansive title. They define the proper contribution: an executable categorical bridge, a recovery certificate, an explicit map of theorem-bearing and heuristic components, and conditional evidence about contamination behavior. Robust-statistics language such as influence and breakdown remains useful for describing the failure modes, but a finite simulation sweep is not by itself a general estimator-level robustness theorem [Huber and Ronchetti, 2009]. The same discipline applies to historical and conceptual scholarship: early work on probability, inverse inference, utility, and collective judgment supplies lineage, not evidence for modern KL, FedGVI, or adversarial-federation claims.

9.5 A falsifiable research program

The next stage should be organized around boundary conditions rather than a larger collection of demonstrations. First, derive a server objective whose minimizer is competitive with `robust_aggregate` while retaining the variational rule's effective-weight control. Recent work on closed-form generalized variational objectives, logarithmic-pool weighting, and robust divergence-weighted federation provides relevant mathematical constraints [Nguyen and Westerhout, 2026, Carvalho et al., 2023, Li et al., 2022]. A useful candidate must recover the standard log-linear pool at zero robustness, state which quantity is bounded, and fail visibly when those assumptions are violated.

Second, broaden the primary estimand across attack targets, hidden states, adaptive adversaries, calibration conditions, and model classes. The decisive falsifier is not a lower average score on one new grid; it is failure of the claimed robustness advantage, recovery identity, or stated uncertainty calibration under a pre-registered extension of the data-generating mechanism. Third, promote the point-estimate neural complement to a posterior-parameterized FedGVI study at source-comparable scale, and test whether the client-side bounded-loss behavior survives capacity, optimization, and posterior-family changes. Finally, move from local transport to multi-machine execution with explicit threat models, authentication, failure handling, and privacy claims; none should be inferred from the current bit-identity result.

9.6 Final position

The strongest conclusion is consequently neither that robust belief sharing has been solved nor that the bridge is merely metaphorical. In the declared categorical setting, the standard-Bayes client limits and project log-linear-pool server identity are tested recovery conditions; away from them, server behavior is measurable, regime-dependent, and partitioned into theorem-bearing, objective-backed, and heuristic claims. That is a modest result, but it is a useful one: it supplies a reproducible starting point from which stronger theorems, broader state spaces, independent implementations, and real federated deployments can be judged without losing the baseline they are meant to extend.

10 Reproducibility: execution record and recovery checks

This section is a machine-verifiable reproducibility certificate. Every value below is computed by the analysis pipeline and injected at render time, establishing a chain of custody from configuration through code to publication. The discipline is the one that gates this project's CI: every prose number is a generated token, every token is emitted by one generator function, and any drift between narrative and computed result fails the build before a green PDF exists [Peng, 2011].

10.1 Determinism contract for seeded scientific results

Reproducing every reported number requires only two recorded inputs: the global seed pinned here and the software environment fingerprinted in the next subsection. The determinism contract fixes the first — it states exactly what is held constant, what is asserted to machine tolerance, and what is deliberately not claimed byte-identical.

- Global seed: 0, threaded through every `np.random.default_rng(seed)`; the global `np.random` state is never used.
- Recovery identities use exact or machine-tolerance assertions; seeded study reports are regression-tested for repeatability under the recorded software environment. Rendered PDF/HTML/slide containers are validated as fresh publication products but are not claimed byte-identical across toolchain versions.

- No mocks anywhere: every test is a genuine computation on small categorical distributions or a seeded simulation under this repository's explicit no-mocks policy.

10.2 Environment fingerprint for the reported run

The second reproduction input is the exact toolchain. Every field below is captured by the successful full test-and-coverage receipt before final variable generation rather than transcribed by hand. The receipt is bound to the source, tests, manuscript, source-owned documentation, release metadata, ISC tree, dependency lock, and fresh analysis receipt. It rejects any pre/post-suite drift in that boundary, so a reader matching this environment and the seed above can reproduce the seeded results; the config hash lets them confirm they are running the configuration from which this manuscript was rendered.

Table 8: Software and configuration fingerprint for the hydrated manuscript. The build epoch is derived from `SOURCE_DATE_EPOCH`; an unreleased build records an explicit omitted sentinel rather than wall-clock time.

Field	Value
Python	3.13.11
NumPy	2.4.2
SciPy	1.18.0
PyTorch (MLP complement)	2.12.1
Platform	Darwin arm64
Config hash (SHA-256, first 16)	cf4bfe1fbcb7d6ed
Reproducible build epoch (UTC)	2026-08-09T04:56:18Z

The exact environment used for the reported run is recorded in tbl. 8.

10.3 Reader-surface accessibility boundary

The validated HTML manuscript is the canonical accessibility-enhanced reading surface. Its source gate requires a page language and title, a skip link and main landmark, non-empty image alternatives, figure captions, labelled full-size links, unique identifiers, resolved references, and present local assets on every generated page. These deterministic checks are not a claim of WCAG conformance: alternative-text quality, contrast, keyboard behavior, reading order, reflow, mathematics, and assistive-technology behavior still require manual review.

The combined manuscript and slide PDFs are checked structurally, textually, through retained renderer logs, and by raster inspection. They are not claimed to be tagged or PDF/UA-conformant. A future accessible-PDF claim requires a tagged producer, a dedicated conformance report, and screen-reader and reading-order review; `qpdf` structure checks and successful text extraction alone are insufficient.

10.4 Test and coverage evidence for the claim surface

- Acceptance criteria: 259 total, 257 passing.
- Project test suite: 1559 collected cases; the bound successful receipt records zero failed cases. The project no-mocks policy remains a separately executable source contract.
- Line coverage on `src/`: 90.04% (achieved by the bound full gate; $\geq$ 90% line coverage is enforced in CI, while branch coverage is tracked separately in CI).

To regenerate this evidence from a clean checkout, run the project suite under the pinned development environment; the same invocation is the CI gate, so a passing local run and a green build are the same event:

```
uv run --extra dev pytest tests/ \
  --cov=src --cov-fail-under=90
```

For a release-facing hydration, use the receipt-producing wrapper after any required provisional pre-test render, then rerun hydration without its provisional flag:

```
uv run --extra dev python scripts/validate_test_coverage.py
uv run python scripts/z_generate_manuscript_variables.py
```

10.5 Artifact inventory for figures, data, and reports

Table 9: Top-level generated files in `output/figures`, `output/data`, and `output/reports` at token-hydration time. The generated release manifest is the source of truth for the larger recursive publication bundle. Artifacts are regenerable reviewer snapshots and must not be hand-edited.

Category	Count
Figures	61
Data files	6
Reports	32
Total	99

The top-level artifact counts in tbl. 9 complement the recursive, checksum-bearing release manifest.

10.6 Recovery-limit certificate for the client and project-pool corners

The recovery identities are reproducibility checks: the client machinery must return to standard Bayes at its KL/NLL loss limits, and the server heuristic must return to the project's log-linear pool at zero robustness (eq. 7, eq. 6). Under the explicit shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, that pool specializes Eq. 7's message-combination term; it does not reproduce the complete source protocol. These deviations are computed on every build:

- `robust_aggregate(robustness=0)` versus `log_linear_pool` (eq. 6): 0
- `generalized_posterior(KLD, NLL)` versus closed-form Bayes: 5.55e-17
- Rényi divergence versus KL as $\alpha \to 1$: 0
- β-loss versus NLL as $\beta \to 0$: 0
- rcce versus NLL as $q_{\text{loss}} \to 0$: 0

Any drift in these limits beyond machine precision would mean the robust generalization no longer contains its standard-Bayes client limit and project-local log-linear-pool server limit (sec. 7.1) — and would fail the core test suite before this certificate could render. The certificate covers the recovery identity and the client-side result under the cited source theorem's matching assumptions only; the server-side `robust_aggregate` heuristic is certified here for its recovery limit alone, not for any bounded-influence property (sec. 8.13).

11 Supplement: variational aggregation objective and weight control

This supplement gives the full derivation behind sec. 5.8.2: the server-side aggregator `robust_aggregate` is a heuristic, and a single change of divergence direction turns it into block-coordinate descent on a stated free energy with a derived, redescending effective-weight update. We work throughout with categorical local posteriors $q_n(s)$ over the shared latent factor, base weights $w_n > 0$, robustness $c > 0$, and the consensus q on the probability simplex. Let $\lambda > 0$ be the entropy-weight coefficient, with the current default at $\lambda = 1.0$.

11.1 Why the sharp heuristic is not yet variational

The sharp server rule is empirically strong, but this repository has not established a variational certificate for it. The heuristic of eq. 7 alternates a *reverse-KL* weight update $a_n \leftarrow w_n \exp(-c\,\mathrm{KL}(q_n \,\|\, q))$ with the log-linear consensus $q \leftarrow \mathrm{softmax}(\sum_n a_n \log q_n)$. For the natural direct objective $\sum_n a_n \mathrm{KL}(q_n \,\|\, q) + \frac{1}{c}\mathrm{KL}_{\text{gen}}(a \,\|\, w)$, the reverse-KL rule is the a-minimizer, but the q-minimizer is the *arithmetic* (linear) pool $q \propto \sum_n a_n q_n$, not the log-linear pool. The executable orientation witness confirms this finite-simplex mismatch. The following proposition goes further, while retaining a deliberately narrow scope.

$$Q(a; s) \;=\; \mathrm{softmax}\left(\sum_n a_n \log s_n\right) \tag{20}$$

The proposed raw q-block is the map in eq. 20.

$$F(q, a; s, w) \;=\; \sum_n a_n \,\mathrm{KL}(q \,\|\, s_n) + R(a, w) + G(q). \tag{21}$$

Proposition 9 (Scoped separable raw-log-pool no-go)**.** *For categorical state dimension $K \geq 2$, no objective of the declared displayed form with G continuously differentiable, R independent of q and the s_n, and G independent of a and the s_n, has $Q(a; s)$ as its q-coordinate minimizer for every interior raw $a \in \mathbb{R}^N_{>0}$ and every interior local-posterior collection s. Consequently, this objective class cannot realize both block maps of the implemented raw-weight heuristic.*

Proof sketch. Fix any non-uniform interior q, a positive scalar α, and one local posterior constructed in eq. 22:

$$s_i^{(\alpha)} \;=\; \frac{q_i^{1/\alpha}}{\sum_j q_j^{1/\alpha}}. \tag{22}$$

Then $Q(\alpha; s^{(\alpha)}) = q$. Writing Π for projection onto the tangent space of the simplex, first-order stationarity of eq. 21 at that same q requires $\Pi[(\alpha - 1)\log q + \nabla G(q)] = 0$. The unit-scale construction forces $\Pi \nabla G(q) = 0$; any different positive scale then forces $\Pi \log q = 0$, contradicting the non-uniform choice of q. The executable witness records both exact log-pool identities and the nonzero tangential contradiction.

A companion witness also blocks the obvious normalized-weight escape within the same natural data-term class: two interior consensuses yield the same normalized reverse-KL weights but different forward-KL data-term differences, so a q-independent differentiable $R(a, w)$ cannot satisfy both simplex-stationarity equations. The implementation itself uses raw effective weights, so this companion is a scope check rather than a description of the production update.

Table 10: Formal MAJ-1 witness inventory. These are deterministic finite-simplex proof artifacts, not empirical estimates; no resampling interval or deployment claim is implied.

Formal artifact	Executable source	Scope
Raw-log-pool contradiction	`server_theory.py`: raw witness	The declared raw q-block map over every interior input
Normalized-weight companion	`server_theory.py`: normalized witness	The normalized reparameterization of the same forward-KL data-term class
Typed source report	report `formal_no_go` field	Deterministic witness metadata, separate from the empirical attack grid

Table tbl. 10 records the deterministic implementation surfaces that bind this scoped result to the typed analysis report.

The proposition does **not** say that no objective of any kind exists. It does not exclude nonseparable q–a couplings, source-dependent terms, non-differentiable constructions, or objectives that encode selected fixed points without reproducing the update blocks for all interior inputs. Thus sec. 5.8 retains the heuristic label and claims only the recovery limit, the scoped negative result above, and conditional empirical behavior — never a bounded-influence property or an objective-backed status.

11.2 Aggregation free energy and its block minimizers

Definition 10 (Aggregation free energy)**.** For $c > 0$, $\lambda > 0$, consensus q, and effective weights $a = (a_n)$, $a_n \geq 0$, define $F_\lambda(q, a)$ as in (8) with the forward cross-entropy $\mathrm{CE}(q, q_n) = -\sum_i q_i \log q_{n,i}$, the consensus entropy $H(q)$, and the generalized KL $\mathrm{KL}_{\mathrm{gen}}(a \,\|\, w) = \sum_n [a_n \log(a_n/w_n) - a_n + w_n]$.

The q-block. For $\lambda > 0$, fixing a, the q-dependent part of F_λ is $\sum_n a_n \mathrm{CE}(q, q_n) - \lambda H(q) = \sum_i q_i [\lambda \log q_i - \sum_n a_n \log q_{n,i}]$. Adding a Lagrange multiplier for $\sum_i q_i = 1$ and differentiating gives $\lambda \log q_i + \lambda - \sum_n a_n \log q_{n,i} + \mu = 0$, i.e.

$$q_i \;\propto\; \exp\Big(\tfrac{1}{\lambda} \textstyle\sum_n a_n \log q_{n,i}\Big) \;=\; \mathrm{softmax}\Big(\tfrac{1}{\lambda} \textstyle\sum_n a_n \log q_n\Big)_i, \tag{23}$$

the product of the weighted experts — the consensus update of eq. 9. At the default $\lambda = 1.0$, the $-H(q)$ term sharpens the weighted geometric mean into the product-of-experts (the entropy bonus that makes the project's log-linear pool a product rather than a geometric average).

The a-block. Fixing q, $\partial F/\partial a_n = \mathrm{CE}(q, q_n) + \frac{1}{c} \log(a_n/w_n) = 0$, so

$$a_n \;=\; w_n \exp(-c\,\mathrm{CE}(q, q_n)), \tag{24}$$

the weight update of eq. 9. Because $\mathrm{CE}(q, q_n) = H(q) + \mathrm{KL}(q \,\|\, q_n)$, the forward direction $\mathrm{KL}(q \,\|\, q_n)$ — not the heuristic's reverse $\mathrm{KL}(q_n \,\|\, q)$ — is the one consistent with the consensus update.

Each block update is the *exact* minimizer of its block, so alternating them is block-coordinate descent: F is non-increasing at every half-step. When the iterates converge, their fixed point is coordinatewise stationary.

The implementation keeps numerical failure handling outside that theorem. If finite-precision underflow collapses all effective weights, it records a fallback event, substitutes the declared base weights to return a valid probability vector, and does not certify the substituted trajectory as converged. Such a trace is diagnostic evidence about the solver boundary, not an instance of the exact block-descent result.

11.3 Formal properties of the conservative server rule

Theorem 11 (Variational aggregation: descent, recovery, and effective-weight bound)**.** *Let $c > 0$ and $\lambda > 0$. Each alternating update in (23)–(24) never increases F. Any converged fixed point is coordinatewise stationary. As $c \to 0$ the generalized-KL penalty forces $a_n \to w_n$ and the consensus is the tempered log-linear pool; at the default $\lambda = 1.0$ it is the log-linear pool (6) exactly, so the variational aggregator shares the project log-linear-pool corner of (7). Under the qualified bridge of Section 5.8, this is only the categorical message-combination specialization, not the complete source protocol. Finally, the effective weights satisfy $a_n = w_n \exp(-c\,\mathrm{CE}(q, q_n)) \leq w_n$ with $a_n \to 0$ as $\mathrm{KL}(q \,\|\, q_n) \to \infty$. Thus the raw effective-weight update is bounded and redescending relative to the realized consensus. This statement does not by itself establish a bounded influence function or finite gross-error sensitivity for the normalized consensus estimator.*

The objective F is biconvex (each block convex, the coupling $\sum_n a_n \mathrm{CE}(q, q_n)$ bilinear), so the result concerns monotone block updates and converged coordinatewise fixed points, not guaranteed convergence to or certification of a global minimum.

The effective-weight regime, and why multi-start matters. The weight bound $a_n \leq w_n$ is unconditional ($\mathrm{CE}(q, q_n) \geq 0$ always). The *collapse* $a_n \to 0$ is driven by the agent's divergence *from the realized consensus* $\mathrm{KL}(q \,\|\, q_n)$, and the consensus itself depends on the weights. Because F is biconvex, this couples into a subtlety an adversarial review of this work surfaced: a *near-one-hot* saboteur (contamination rate $\to 1$) already captures the product-of-experts, so a descent seeded *at* that pool stays in a consensus-capture basin (high F) where the saboteur keeps its weight — even against an honest majority. The repair is to search the stated objective more carefully: `variational_aggregate` runs **multi-start** block-coordinate descent (the pool, the uniform belief, and the arithmetic-mean seeds) and returns the lowest-observed-F converged candidate. In the configured colony, the uniform/arithmetic seeds reach a lower-F *vetoing* basin, so the saboteur is suppressed even at the simplex vertex (pinned qualitatively by the near-vertex multi-start test) — the 267.1× suppression of fig. 15 is measured across the swept contamination grid, whose most extreme point sits just below rate 1. What remains fundamental to *every* robust fusion rule, and is not claimed away: with no honest majority — a colony split with no anchoring plurality — there is no truth to recover. The observed suppression is conditional on the tested colonies and the fixed point selected by a finite multi-start heuristic.

fig. 17 makes the capture and the escape concrete on a near-vertex colony: the single (log-linear-pool) start settles at $F = 1.3092$ (the capture basin, where the saboteur keeps its weight), while the multi-start descent reaches the genuinely lower $F = -0.2305$ vetoing basin — a gap of 1.5397 nats that is exactly the difference between trusting the natural seed and solving the stated objective properly.

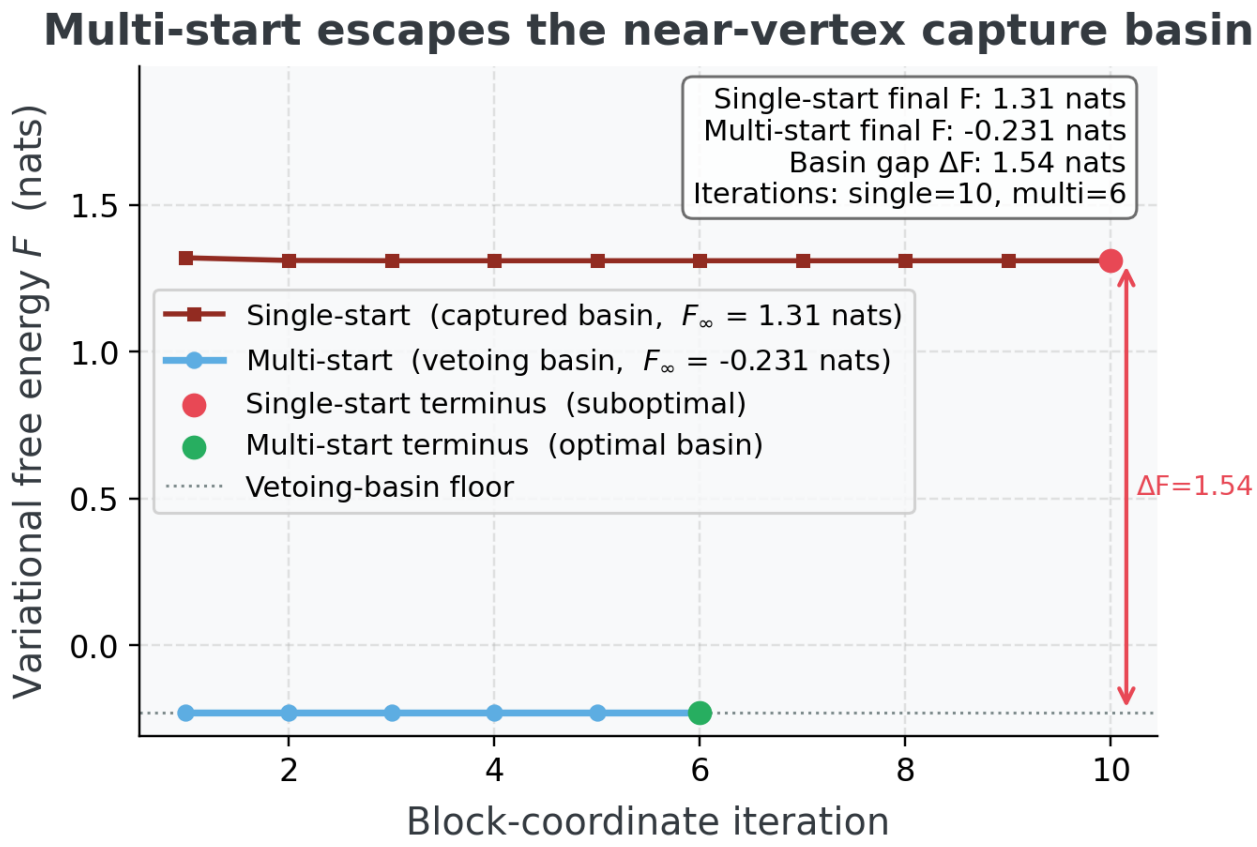


Figure 17: Variational free-energy descent on a near-vertex adversarial colony. Source relation: original project objective-descent diagnostic; estimand: free energy F in nats by iteration; uncertainty: none for deterministic seeded runs. The figure compares the single (log-linear-pool) start versus the multi-start descent. The x-axis is the block-coordinate iteration; the y-axis is the free energy F in nats. The single-start trajectory settles in the high-F capture basin ($F = 1.3092$, the saboteur retains weight); the multi-start trajectory reaches the lower-F vetoing basin ($F = -0.2305$), a gap of 1.5397 nats. Deterministic seeded runs, so no error band.

11.4 Numerical witnesses for descent and influence bounds

The analysis pipeline runs `variational_aggregate` at robustness $c = 1.50$ on a contaminated colony and records the free energy after each iteration. The descent falls from $F = 3.2458$ to $F = 2.3780$ (a monotone drop of 0.8678 over 11 iterations, converged: Yes); the largest single-step *increase* is 8.88×10^{-16}, machine zero — the monotonicity of the theorem, witnessed numerically and drawn in fig. 14.

For the effective-weight diagnostic, one agent is drifted from healthy toward a confident-wrong delta and its normalized influence is read at each drift. Clean, it carries 0.143 of the pool; at the most extreme swept drift it carries below 0.001 — a factor of 267.1 (computed from the unrounded influences, not the display-rounded values above) below the fixed 0.143 the naive pool would still grant it (fig. 15). This makes the redescending normalized-weight behavior visible on the tested path; it is not an estimator-level B-robustness proof.

11.5 Tempered aggregation family for the accuracy-guarantee trade

The aggregator of sec. 11.2 fixes the entropy term at unit weight. Relaxing that single coefficient generates a one-parameter *tempered* family. Introduce an entropy weight $\lambda > 0$ — the **inverse temperature is** $1/\lambda$ — and minimize

$$F_\lambda(q, a) \;=\; \sum_n a_n \,\mathrm{CE}(q, q_n) \;-\; \lambda\, H(q) \;+\; \tfrac{1}{c}\,\mathrm{KL}_{\mathrm{gen}}(a \,\|\, w). \tag{25}$$

Repeating the q-block derivation of eq. 23 with the entropy scaled by λ leaves the a-block **untouched** and tempers only the consensus update:

$$q \;\propto\; \exp\!\Big(\tfrac{1}{\lambda}\textstyle\sum_n a_n \log q_n\Big), \qquad a_n \;=\; w_n \,\exp(-c\,\mathrm{CE}(q, q_n)). \tag{26}$$

The $\lambda \downarrow 0$ endpoint is separately implemented as a deterministic tied-argmax rule; it is not obtained by substituting $\lambda = 0$ into eq. 25 or eq. 26.

The weight update is **independent of** λ: the bound $a_n \le w_n$ with collapse $a_n \to 0$ as $\mathrm{KL}(q \,\|\, q_n) \to \infty$ is unchanged, so the **raw effective-weight bound of sec. 11.3 holds for every** $\lambda > 0$. At $\lambda = 1.0$ the temperature is unity and eq. 26 is **identical** to the current axis-3 aggregator eq. 9 — the default is bit-identical, not merely close. The $c \to 0$ recovery of sec. 11.3 generalizes to the *tempered* log-linear pool $q \propto \exp(\frac{1}{\lambda}\sum_n w_n \log q_n)$; at $\lambda = 1.0$ this is exactly $\mathrm{softmax}(\sum_n w_n \log q_n)$ — the project's **log-linear pool** eq. 6. Under the shared-support, posterior-log-potential, and fixed-weight assumptions of sec. 5.8, that pool is a categorical specialization of Friston Eq. 7's message-combination term, not a reconstruction of the complete source protocol. Positive-temperature members away from that default are tempered pools, not Friston Eq. 7 itself; the project recovery checks, including ISC-10, remain project-local.

A small empirical sweep over $\lambda \in \{0.1, 0.2, 0.3, 0.5, 0.7, 1\}$ on 10 contaminated colonies (5 agents, 2 adversarial) asks whether a single λ^* makes the conservative aggregator narrow the gap to the sharp robust_aggregate point-accuracy. The closest observed weight is $\lambda^* = 0.3$ with an accuracy gap of 0.0008. **A lambda* narrows the tested accuracy gap while preserving the stated weight bound on this grid.** If no λ closes that gap while preserving the derived weight update, the result is the conservatism trade-off of sec. 8.13, not a defect to hide.

12 Supplement: extended methods for scoped generalization

This supplement documents three method extensions that broaden the toolkit without changing the categorical federated claims of the main text. Each answers a "does it generalize?" question raised by a specific main-text result and each connects back to it: the additional contamination models, gallery, and onset sweep stress-test the robustness verdict of sec. 7.5 beyond its single confident-wrong mechanism; the Gaussian divergence bridge points toward the continuous-state direction of sec. 8.22; and greedy multi-hypothesis reduction extends the emergence study of sec. 7.4 from one pruned state to a family. Each is tested and isolated; none participates in the headline robustness verdict, so the main-text claims stand or fall without them.

12.1 Continuous-state divergence bridge for Gaussian beliefs

Every robustness claim in the main text is discrete-categorical, matching Friston's worked example. To show the divergence family carries over to the Gaussian beliefs a continuous-state active-inference extension would use, `divergences.py` adds closed forms for 1-D Gaussians: the Kullback–Leibler divergence

$$\mathrm{KL}(\mathcal{N}(\mu_q, \sigma_q^2) \,\|\, \mathcal{N}(\mu_p, \sigma_p^2)) = \tfrac{1}{2}\left[\tfrac{\sigma_q^2}{\sigma_p^2} + \tfrac{(\mu_p - \mu_q)^2}{\sigma_p^2} - 1 + \log \tfrac{\sigma_p^2}{\sigma_q^2}\right], \tag{27}$$

and the α-Rényi divergence with interpolated variance $\sigma_\alpha^2 = \alpha\sigma_p^2 + (1-\alpha)\sigma_q^2$,

$$D_\alpha(\mathcal{N}_q \,\|\, \mathcal{N}_p) = \frac{\alpha(\mu_q - \mu_p)^2}{2\sigma_\alpha^2} - \frac{1}{2(\alpha - 1)} \log \frac{\sigma_\alpha^2}{\sigma_q^{2(1-\alpha)}\sigma_p^{2\alpha}}. \tag{28}$$

As in the categorical case (eq. 3 and Lemma 3), eq. 28 recovers eq. 27 in the $\alpha \to 1$ limit, and the closed form is returned exactly inside a small band around one. These functions are **out of scope** for the federated experiments — they are wired into no aggregation rule or sweep — and exist purely as the explicitly-scoped bridge toward continuous active inference.

12.2 Additional contamination models for the robustness surface

`contamination.py` extends the confident-wrong, label-noise, and uniform models with two mechanisms that probe different attack surfaces, both honoring the identity anchor (rate zero returns the belief unchanged):

- **Byzantine targeted** — a *multiplicative* log-odds tilt toward an adversary-chosen state, $s' \propto s \cdot \exp(\mathrm{rate} \cdot \mathrm{tilt} \cdot e_{\mathrm{target}})$. Unlike the additive convex mixes, the corruption compounds with the belief's own shape, so the non-target states keep their relative order — the canonical targeted poisoning of a product-of-experts pool.
- **Drift** — a *slowly-moving* bias that grows linearly across communication rounds via a phase $\phi = \mathrm{round}/(\mathrm{rounds} - 1)$, so the first round is clean and the bias creeps in. This is the stealthy sentinel whose miscalibration only becomes confident late, defeating any one-shot screen.

Both are exercised in the contamination tests; the headline sweep continues to use the confident-wrong model so the verdict is comparable to the main text.

12.2.1 Contamination gallery by corruption mechanism

To check the robust-beats-naive result is not an artifact of the single confident-wrong mechanism — and not of a single lucky seed — `experiments.run_contamination_gallery` re-runs the paired comparison ($n = 24$ trials across 64 independent seeds, contamination strength 0.60) under every model. For each mechanism it selects one robust method by pooled mean consensus accuracy **for descriptive gallery display only**, then reports that displayed member's robust-minus-naive difference, 95% seed bootstrap interval, and *win fraction* — the fraction of seeds in which the displayed member beats naive. This is not the selection-free inferential surface; the all-method review grid below serves that role.

Table 11: Seed-aggregated descriptive display of robust-vs-naive accuracy under each contamination mechanism ($n = 24$ trials $\times$ 64 seeds at strength 0.60). `Reliable` is a display flag for the pooled-selected member: it is `Yes` only when that member beats naive in at least 0.95 of seeds *and* its displayed difference CI excludes zero. This is an across-seed screen, not one lucky seed, a p-value, or selection-free post-selection inference.

Mechanism	Class	Naive	Best robust	Mean diff	95% CI	Win frac.	Reliable
byzantine	directional	0.6306	0.6599 (beta)	0.0293	[0.0124, 0.0472]	0.62	No
confident wrong	directional	0.9836	0.9878 (AR)	0.0043	[0.0040, 0.0046]	1.00	Yes
drift	directional	0.9836	0.9878 (AR)	0.0043	[0.0040, 0.0046]	1.00	Yes
label noise	entropy	0.9982	0.9931 (AR)	-0.0051	[-0.0052, -0.0051]	0.00	No
uniform	entropy	0.9985	0.9947 (AR)	-0.0038	[-0.0038, -0.0038]	0.00	No

The contamination summary tbl. 11 is a descriptive sensitivity screen, deliberately narrower than "robust always wins." The pooled display member has a positive all-seed and interval pattern under confident wrong, drift — the *additive* directional attacks. The full set of directional mechanisms is confident wrong, byzantine, drift; the **byzantine** attack is directional too, but its *multiplicative* log-odds tilt escalates faster: at this strength it sits near a veto cliff where the naive pool is already badly degraded and the displayed robust advantage does not hold across seeds (its win fraction is well below the 0.95 display bar and its difference CI straddles zero), so we do *not* claim it. The **entropy** attacks (label noise, uniform) raise entropy or inject noise without a fixed wrong target, so the product-of-experts is not pulled off the truth and there is nothing to beat — the robust members stay close rather than winning (naive undegraded by entropy attacks: Yes). fig. 18 draws all mechanisms with their win fractions. This is the honest scope of this configured gallery: its displayed members separate from naive under the declared *sustained additive* directional contamination, stay close under the declared entropy attacks, and lose the displayed advantage against the tested multiplicative adversary near the veto regime. These finite cells do not establish the same ordering for every attack strength or world, and they do not turn a pooled display selection into selection-free inference.

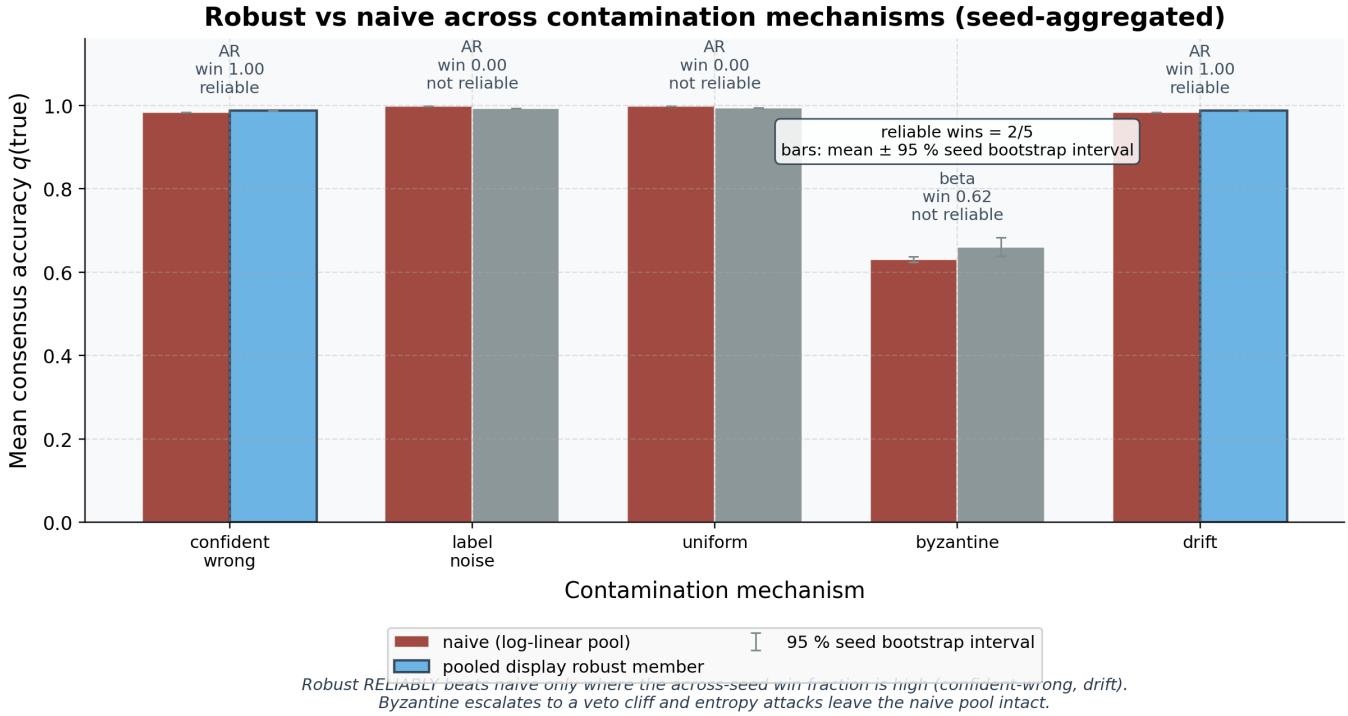


Figure 18: Seed-aggregated mean consensus accuracy. Source relation: original project contamination diagnostic; estimand: true-state accuracy fraction by attack mechanism; uncertainty: the bars show 95% seed-level bootstrap confidence intervals for the pooled-selected display member, while the adjacent table reports its conditional paired difference interval. q(true state) for the naive log-linear pool versus the robust method selected once by pooled mean under each contamination mechanism ($n = 24$ trials $\times$ 64 seeds at strength 0.60). The x-axis is the contamination mechanism; the y-axis is mean consensus accuracy. Each group has two bars: naive log-linear pooling and the pooled display member for that mechanism. The robust bar is drawn in full color only where the across-seed win fraction (annotated above the group) clears the 0.95 display bar — confident wrong, drift; the byzantine mechanism and entropy attacks are muted because they do not clear that descriptive screen. The in-figure summary gives the display-flag count across mechanisms and reminds readers that the labels are win fractions, not p-values. The bars are means over 64 seeds; the selected method is shown above each bar. This is a descriptive pooled-selection graphic, not selection-free post-selection inference; the all-method review grid supplies the latter surface.

12.2.2 Robustness onset by corruption mechanism

The gallery fixes one contamination strength; `experiments.run_robustness_onset` maps the *rate dependence* ($n = 24$ trials $\times$ 64 seeds per rate). For each directional mechanism it reports the **descriptive onset rate** — the smallest rate at which the pooled-selected display member's win fraction reaches 0.95 — and that member's versus naive accuracy at the worst (highest) swept rate. These display summaries are not selection-free inference; the all-method review grid is the inferential surface:

Table 12: Per-mechanism descriptive onset and worst-rate accuracy ($n = 24$ trials $\times$ 64 seeds). The onset rate is where the pooled display method reaches the displayed win-fraction rule (win fraction >= 0.95); it is not a per-seed selection, selection-free inferential result, or universal crossover claim.

Mechanism	Onset rate	Naive @ worst	Robust @ worst	Robust method @ worst
byzantine	0.4	0.0165	0.0000	beta
confident wrong	0.6	0.6676	0.7772	beta
drift	0.6	0.6676	0.7772	beta

The mechanism-specific onset thresholds are collected in tbl. 12.

The rate dependence sharpens the gallery's snapshot, and fig. 19 draws it. The additive confident-wrong and drift attacks degrade the naive pool gradually; past their onset rate the robust member stays above it through to the worst rate. The multiplicative byzantine attack is qualitatively different: it opens an *early* robustness window — robust overtakes at a lower onset rate — but then escalates to the veto cliff where naive and robust both collapse, so its worst-rate accuracy is near zero for both. This is the rate-resolved form of the honest verdict: robustness is sustained against additive directional contamination and only transient against a multiplicative one.

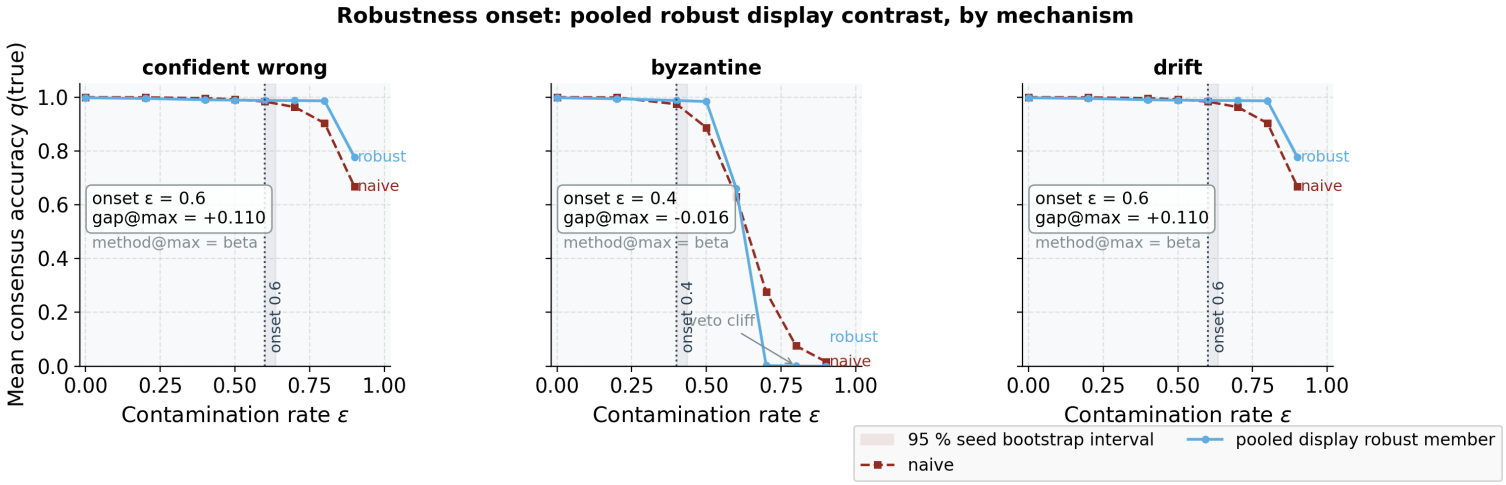


Figure 19: Naive (dashed) versus the pooled display method. Source relation: original project robustness-onset diagnostic; estimand: mean consensus accuracy fraction by attack rate; uncertainty: shaded 95% seed-level bootstrap confidence intervals conditional on the pooled-selected display member. Mean consensus accuracy (solid, robust method selected once by pooled mean across seeds at each rate; dashed, naive) as the contamination rate rises, one panel per directional mechanism ($n = 24$ trials $\times$ 64 seeds per rate). The x-axis is the contamination rate; the y-axis is mean consensus accuracy. The dotted vertical line marks the descriptive onset rate (pooled robust win fraction >= 0.95), and each panel's inset reports that onset plus the final pooled robust-minus-naive gap at the largest swept rate. Confident-wrong and drift show a sustained displayed contrast past onset; byzantine shows a transient display window before both aggregators lose consensus accuracy at the highest corruption rates. The plotted values are seed-aggregated means with shaded bootstrap intervals; the companion table carries the displayed onset, worst-rate values, and selected method. This pooled-selection display is not selection-free post-selection inference; the all-method review grid supplies that inferential surface.

12.2.3 Conditional world and attack-geometry grid

The finite MAJ-1 characterization is now extended across 40 preregistered world/scenario cells: two hidden-state locations, two observability levels, five attack mechanisms, and two adversarial weight settings. The independent unit is the seeded world/scenario row; each cell averages 24 nested trials over 64 seeds before the matched contrast is formed. The primary estimand is naive true-state error minus robust true-state error, so a positive value means the robust heuristic assigns more true-state mass in that finite cell. The robustness-zero control is pass, and the report remains explicitly labelled `conditional_finite_grid`. The resulting conditional surface is shown in fig. 20.

12.2.4 Source-bound robustness review grid

The red-team review adds a bounded, selection-free stress surface that joins the existing conditional-world cells to the existing directional rate profiles. It uses 160 deterministic seed replicates and 24 trials nested within each seed/cell. The finite attack union is clean, confident wrong, permutation,

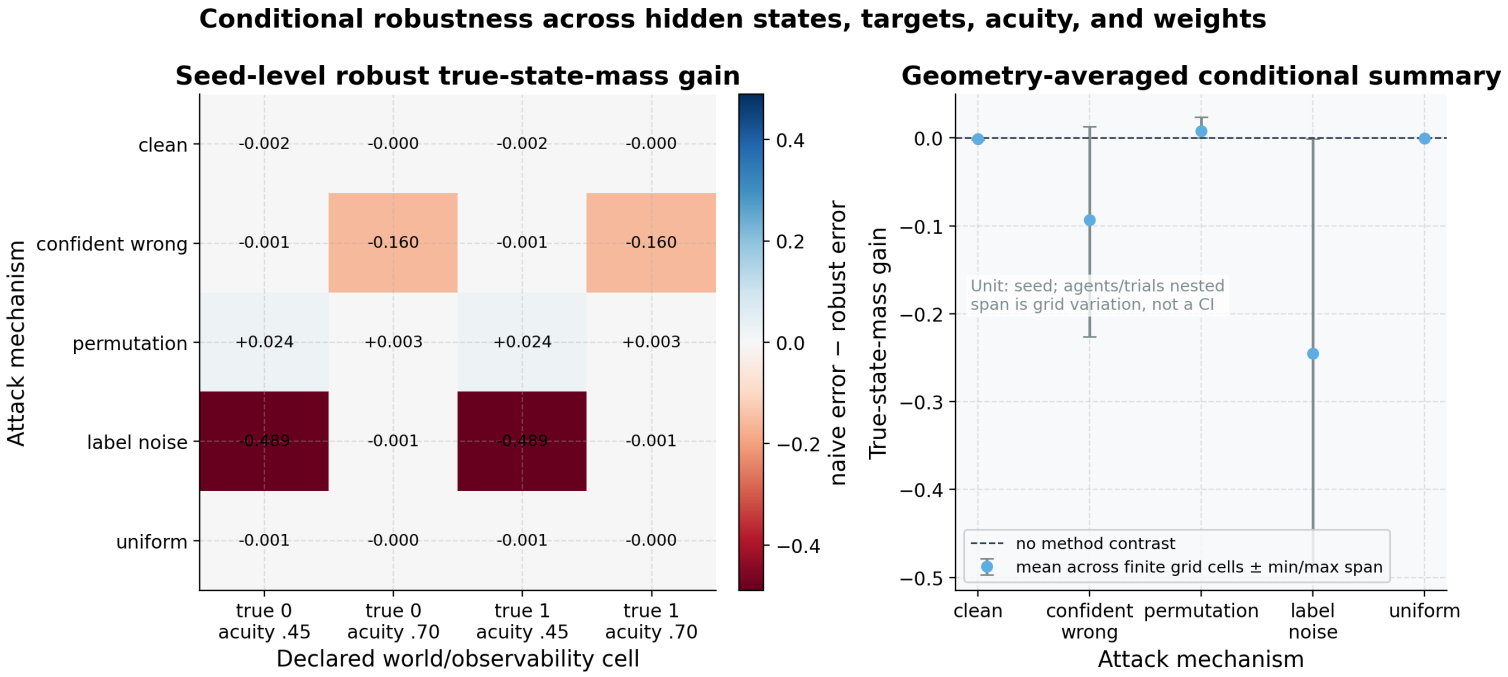


Figure 20: Conditional-world robustness grid. Source relation: original project finite-grid generalization of the MAJ-1 characterization; estimand: naive true-state error minus robust true-state error; uncertainty: each heatmap cell is a seed-level mean with a 95% seed bootstrap interval in the source report, while the right panel shows finite-grid min/max span rather than a confidence interval; independent unit: seeded world/scenario row. The x-axis is the declared hidden-state and observability cell; the y-axis is the attack mechanism. The left panel varies hidden state and observability across columns and attack mechanism across rows; the right panel summarizes the finite-grid span by attack. Positive values favour robust true-state mass, negative values favour naive pooling, and zero is the recovery/no-contrast reference. This is conditional evidence over a declared finite grid, not a theorem, breakdown bound, or universal attack result.

byzantine, drift, label noise, uniform; the rate-resolved directional mechanisms are confident wrong, byzantine, drift, with entropy controls uniform, label noise. The registered rate set is $\{0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. The independent unit is seed within a declared scenario or rate cell, and the nesting rule is: n_trials nested within each seed/cell; no trial is promoted to an independent world. This is a source-bound simulation review, not an external-data benchmark or a claim that cells sharing design structure are independent.

The payload is explicitly selection-free source payload; every configured non-KLD method is reported at every directional rate and no winner is used for inference and the statistics surface is selection-free. It reports seed-level contrasts, paired Wilcoxon/rank-biserial results, percentile bootstrap intervals, MCSE, an observed-effect MDE, and BH-adjusted rate families. BH ownership is BH is applied within each attack-mechanism × method rate family; cells sharing design structure are not treated as independent families. The precision plan targets maximum MCSE 0.0100 and observed maximum MCSE 0.0066 across 96. Every configured robust method is retained as a rate-profile curve and inferential member; no method is selected per seed, rate, or pooled mean for this review grid. The all-method display does not close the open calibration or server-theory questions.

The rendered diagnostic is shown in fig. 21.

12.2.5 Proper scores and calibration controls

Argmax accuracy does not distinguish a cautious posterior from an overconfident one. The scoring extension therefore pre-registers the paired seed-level categorical log-score difference between naive and robust consensus beliefs as its primary belief-quality estimand. The Brier score and equal-width expected calibration error are secondary diagnostics; higher log score and lower Brier/ECE are better. Agents and trials remain nested within the seed.

The report also includes three negative controls: an oracle, a uniform belief, and a confidently wrong belief. Their expected score ordering is checked before any method contrast is interpreted: oracle > uniform > confidently wrong under the clipped log score. The control ordering gate is `pass`, and the confidently-wrong-versus-uniform control is `pass`, using 64 independent seeds and 24 nested trials per seed. The diagnostic is shown in fig. 22.

12.3 Greedy multi-hypothesis model reduction beyond the main BMR study

The single-step reduction of sec. 5.4 scores one candidate reduced prior. `bayesian_model_reduction.py` adds `greedy_reduce`, which performs structure learning over a *family* of redundant states: starting from the full prior, it scores pruning each not-yet-pruned state against the current reduced prior, accepts the single prune with the largest positive free-energy gain, and repeats until no remaining prune improves model evidence. Every accepted step has a strictly positive incremental ΔF, so the cumulative evidence is monotone-increasing and the search recovers the sparse generative model the data support — a state with genuine evidence yields $\Delta F < 0$ when pruned and is kept. This is the multi-state analogue of the emergence result of sec. 7.4, and is verified directly in the model-reduction tests.

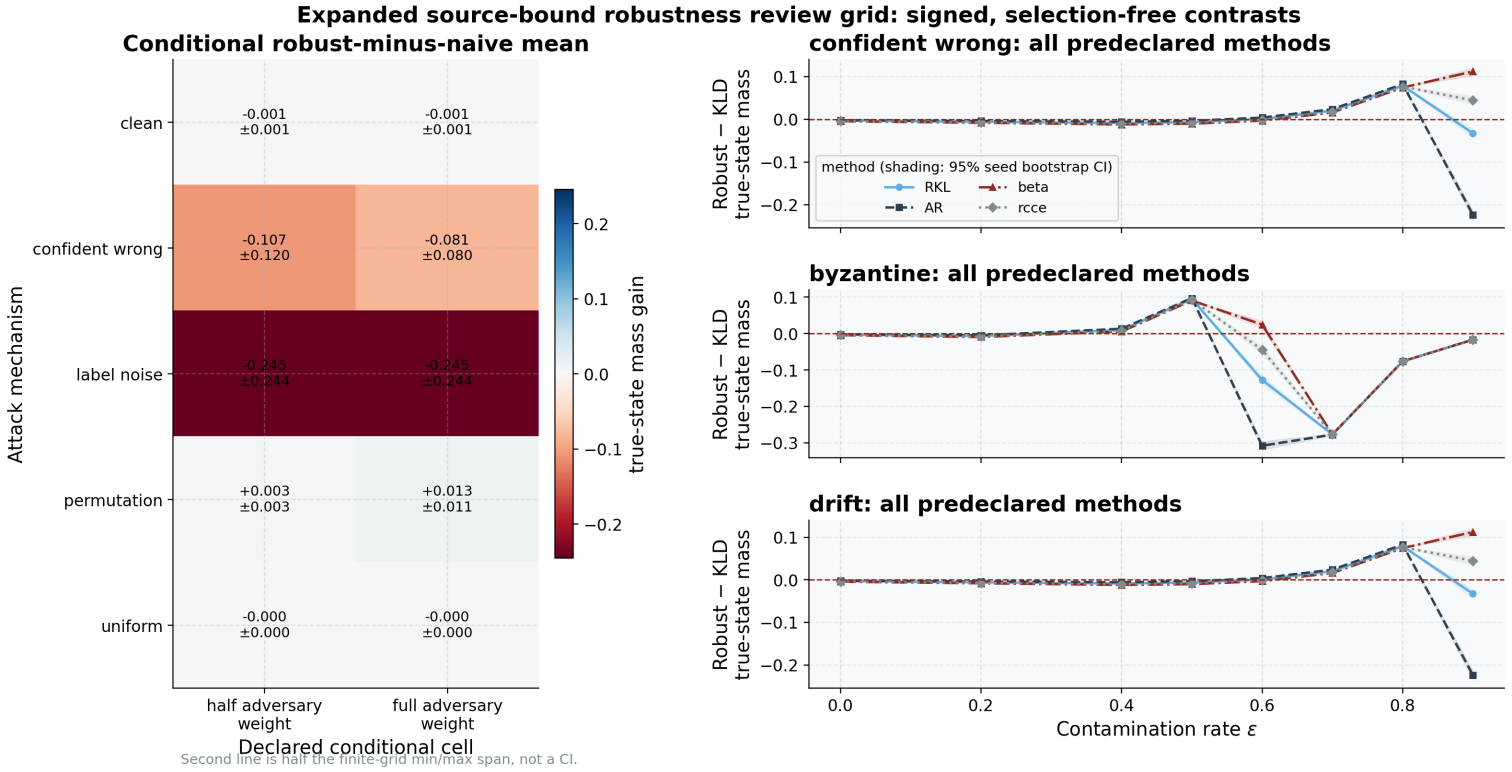


Figure 21: Expanded source-bound robustness review grid. Source relation: original project finite simulation review diagnostic composed from the existing conditional-world and onset mechanisms; estimand: seed-level robust-minus-naive true-state probability-mass contrast; uncertainty: the right-panel shaded bands are percentile bootstrap intervals over independent seeds for every configured robust method, while the second line in each left-panel cell is half the finite-grid min–max span, not a confidence interval; replication unit: configured seed, with trials nested within seed and cell. The x-axis is the declared adversarial-weight setting in the left panel and the contamination rate in the right panel; the y-axis is the seed-level robust-minus-naive true-state mass contrast in both panels. The left panel summarizes conditional attack cells, and the right panel shows every configured directional method's signed rate profile over the registered rates. Positive values favour robust true-state mass, negative values favour naive pooling, and zero is the recovery/no-contrast reference. No method or curve is selected by pooled mean for this grid; all displayed intervals and comparisons are selection-free. This visualization is conditional finite-grid evidence and does not claim a universal winner, breakdown bound, causal effect, or independence across shared design cells.

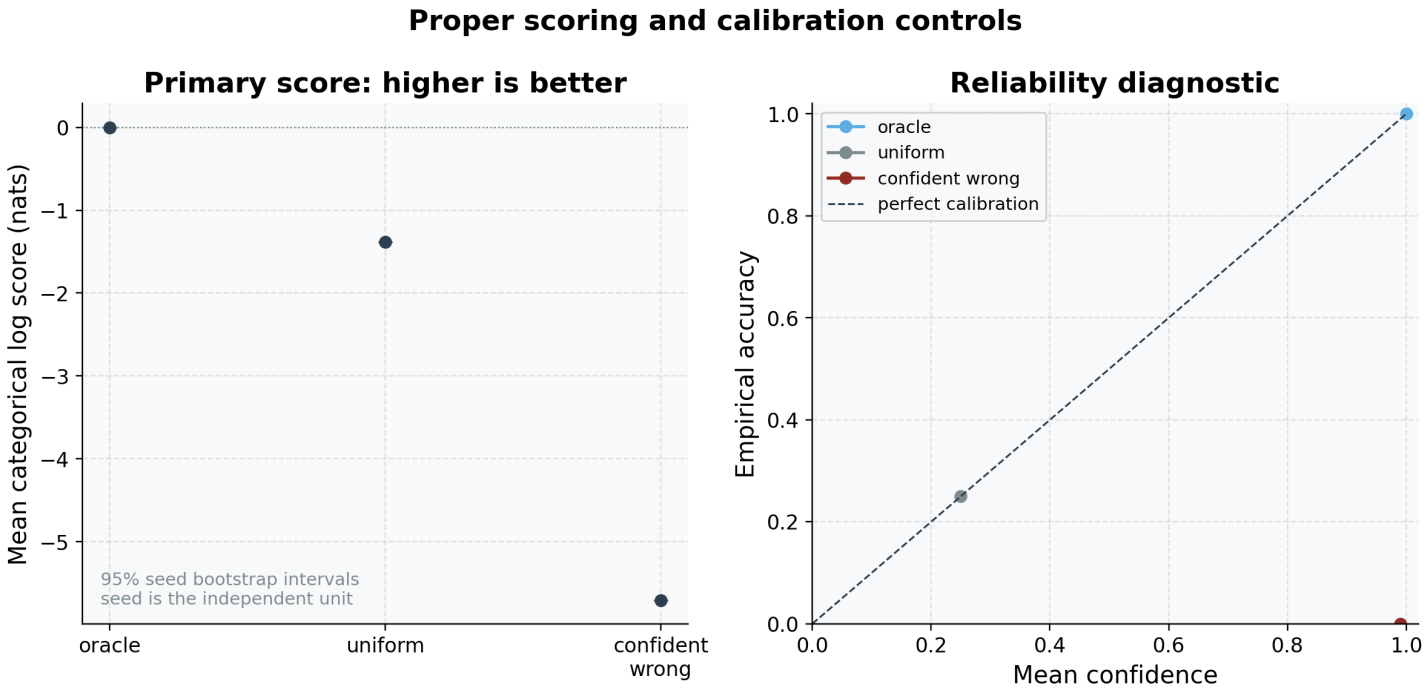


Figure 22: Proper scoring and calibration controls. Source relation: original project belief-quality diagnostic; estimand: categorical log score as the primary measure, with Brier score and reliability error as secondary diagnostics; uncertainty: 95% seed bootstrap confidence intervals for control log scores; independent unit: seed, with trials nested within seed. The x-axis is the control type in the left panel and mean confidence in the right panel; the y-axis is mean categorical log score in the left panel and empirical accuracy in the right panel. The left panel compares oracle, uniform, and confidently-wrong controls on the higher-is-better log-score scale. The right panel plots mean confidence against empirical accuracy for the same controls and a perfect-calibration diagonal. The controls are negative checks on score implementation, not evidence for decision optimality, distribution-shift calibration, or robustness outside the tested finite world.

12.4 Federation transport protocol and bit-identity witness

This supplement answers a concrete question the main text raises but settles elsewhere: when belief sharing is routed through an actual transport channel instead of a direct function call, does the fused consensus change? It specifies that transport — the same single-host interface sec. 8.21 names as the anchor for eventual multi-machine federation — and establishes that, under a lossless round-trip, the answer is no.

Concretely, the transport realizes belief sharing over a real in-memory channel rather than a direct aggregation call. Each worker holds a local posterior q_n over the shared latent factor and serializes it to lossless IEEE-754 float64 bytes using the numpy-lossless-float64 encoding (numpy's native array format), guaranteeing bit-identical round-trip across the transport boundary. A server collects 5 such beliefs, fuses them with the same robust server step at robustness $c = 1.5$, and broadcasts the consensus q back to every contributing worker over its response channel.

Proposition 12 (Federation bit-identity)**.** *When the transport serialization is lossless — an exact IEEE-754 float64 round-trip — the federated consensus equals the in-process aggregation* $q = \text{robust\_aggregate}(\{q_n\}, c)$ **bit-for-bit*. Transport moves bytes, not mathematics, so no precision is lost and no result changes.*

Because the round-trip is exact, this implementation retires the direct in-process serialization caveat. The queue adapter remains a genuine `queue.Queue` transport, `run_multiprocess_round` exercises the same server/worker protocol with one OS worker process per agent on a single machine, and `run_socket_round` exercises the loopback-TCP adapter. The fused result is provably unchanged. Bit-identity verified: True.

The implementation lives in the `federation/` package. The end-to-end and socket transport tests exercise the full round-trip — worker serialization, server aggregation, consensus broadcast, out-of-order arrival, the single-machine process helper, loopback TCP framing, optional HMAC frame integrity, and file-backed digest-verified replay validation — and assert bit-identity against the in-process `robust_aggregate` result. A caller-owned SQLite guard rejects reused round IDs across local process restarts, but does not define a shared multi-host replay domain. This test surface is the API contract that sec. 8.21 identifies as the anchor for future network transport: the aggregation mathematics can remain unchanged, but true multi-machine work still requires cross-host transport, identity-bound mTLS, shared replay state, discovery, restart orchestration, and threat-model validation that this single-host evidence does not supply.

13 Supplemental notation contract

This supplement is the authoritative notation contract for Active Fedference. The methods, formalism, results, figures, report schemas, and API documentation use these meanings even when a source paper uses a different symbol. A symbol is not reused for a different mathematical object merely because the objects are both probability vectors. The implementation names in the final column are the canonical names for new code and reports; old names survive only as warned, parity-tested compatibility adapters.

13.1 Probability objects and generative-model quantities

13.1.1 States, posteriors, and site factors

Symbol	Contractual meaning	Canonical implementation term
s	A hidden categorical state; $s \in \{1, \dots, n_s\}$ is an index, not a distribution.	`state`
o	An observation/outcome index; $o \in \{1, \dots, n_o\}$.	`observation`
$q_n(s)$	Agent n's local posterior over the shared latent state after its local update.	`local_posteriors[n]`
$q(s)$	The server consensus posterior over the shared state.	`global_posterior` / `consensus`
$q_{-n}(s)$	The normalized cavity posterior with agent n's site contribution removed.	`cavity(...)`
$t_n(s)$	Agent n's site/factor term in natural-parameter space.	`site_factor`
$m_n(s)$	Bridge-only source-equation log potential: $m_n(s) = \log q_n(s) + \kappa_n$ with state-constant κ_n. It is not a claim that every source-protocol message is a broadcast posterior.	No project API; notation for the qualified bridge only.

13.1.2 Priors, policies, and POMDP quantities

Symbol	Contractual meaning	Canonical implementation term
$\pi_0(s)$	A prior over hidden states. The subscript distinguishes a prior from a policy.	`prior` / `log_prior`
π	A policy or action sequence; it is not a prior and is bold when needed.	`policy`
$A[o, s] = P(o \mid s)$	Observation likelihood matrix; each state-indexed column is a pmf.	`likelihood`
$B[s', s, u] = P(s' \mid s, u)$	State-transition tensor indexed by next state, current state, and control.	`transition`
$C[o]$	Log-preference over outcomes; $p_C(o) = \text{softmax}(C)[o]$ is the preferred-outcome pmf.	`log_preferences`
$D_0[s]$	Initial hidden-state prior in the POMDP.	`initial_prior`
$q(o \mid \pi)$	Policy-conditional predicted outcome distribution in EFE calculations.	`predicted_outcomes`

The state index s, posterior $q(s)$, prior $\pi_0(s)$, and policy π must not be conflated. In particular, the policy symbol is never used for a prior, and the prior is never called a policy. The uppercase POMDP tensors A, B, C, D_0 are model objects; they are not posterior factors.

For the qualified relation to Friston et al.'s Eq. 7 [Friston et al., 2024], the shared support is finite, every $q_n(s)$ is positive on it, the Eq. 7 softmax input is represented by the bridge-only $m_n(s)$ above, and the declared weights w_n are fixed rather than functions of the emerging consensus. Under exactly those assumptions, additive κ_n constants cancel under softmax and eq. 6 is the categorical posterior-log-potential specialization of the source message-combination term. It neither reconstructs source message construction, cavity/exclusion policy, scheduling, generative factors, nor the complete source protocol.

13.2 Divergences, losses, and scalar controls

13.2.1 Generalized-Bayes and aggregation terms

Symbol	Contractual meaning	Canonical implementation term
$\mathcal{D}(q\|p)$	A regularizing divergence between distributions, such as KL, reverse KL, or α-Rényi.	`divergence`
$L(s; o)$	Loss evaluated at state s for observation o.	`loss_by_state`
$\tau > 0$	Generalized-Bayes learning rate/temperature multiplying the accumulated loss.	`tau` (`learning_rate` is a warned compatibility alias)
w_n	Non-negative raw/base aggregation weight supplied for local posterior q_n.	`base_weights[n]`

Symbol	Contractual meaning	Canonical implementation term
a_n	Raw variational server effective weight before normalization. In the variational rule $0 \le a_n \le w_n$.	`raw_effective_weights[n]`
$\tilde{a}_n = a_n / \sum_m a_m$	Normalized influence weight returned for interpretation and plotting.	`normalized_effective_weights[n]`

The symbols w_n, a_n, and $\tilde{a}_n$ are deliberately distinct. The first is supplied before aggregation, the second is the variational raw server output, and the third is only its normalized influence representation. The server heuristic's reweighting is not a FedGVI client loss and does not inherit a client-side robustness theorem. `robust_aggregate` is a server heuristic with the tested $c = 0$ recovery identity. `variational_aggregate` owns the explicit finite-simplex objective and the raw-weight bound; that bound is not an estimator-level B-robustness theorem.

13.2.2 Robustness, divergences, and loss controls

Symbol	Contractual meaning	Canonical implementation term
$c \ge 0$	Server-side robustness coefficient used by the divergence-reweighting rule.	`robustness`
$\lambda > 0$	Entropy weight in the variational server objective and its coordinate updates. The $\lambda \downarrow 0$ endpoint is a separate deterministic tied-argmax rule.	`entropy_weight`
$\alpha > 0$	Rényi divergence order.	`alpha`
$\beta \ge 0$	Density-power loss parameter.	`beta`
$q_{\text{loss}} > 0$	Robust categorical cross-entropy parameter in $L_{q_{\text{loss}}}$. The $q_{\text{loss}} \downarrow 0$ NLL limit is handled separately, and the subscript prevents collision with posterior $q(s)$.	`q_loss`
$\rho \in [0, 1]$	Contamination strength/rate in a declared attack mechanism.	`contamination_rate` / `rate`

For the objective-backed server rule, the complete scalar-control contract is defined for $c > 0$ and $\lambda > 0$:

$$\begin{aligned} F_\lambda(q, a) &= \sum_n a_n \,\mathrm{CE}(q, q_n) - \lambda H(q) + \frac{1}{c}\,\mathrm{KL}_{\text{gen}}(a \| w), \\ q &\propto \exp\left(\frac{1}{\lambda}\sum_n a_n \log q_n\right), \\ a_n &= w_n \exp[-c\,\mathrm{CE}(q, q_n)]. \end{aligned} \tag{29}$$

The implementation uses $\lambda = 1.0$ by default; the `entropy_weight` argument exposes the stated tempered family. The $c = 0$ branch is handled as a recovery limit outside the $c > 0$ objective; at $\lambda = 1.0$ it is the exact project log-linear pool. The $\lambda \downarrow 0$ endpoint is separately implemented as a deterministic tied-argmax rule and is not obtained by substituting $\lambda = 0$ into the displayed objective or update.

13.3 Cavity and factor algebra

For a positive-support global posterior and site term, the cavity operation is defined in log space and then normalized:

$$q_{-n}(s) = \frac{q(s)/t_n(s)}{\sum_{s'} q(s')/t_n(s')} = \mathrm{softmax}(\log q(s) - \log t_n(s))\,. \tag{30}$$

The corresponding factor replacement is

$$\log t_n^{\text{new}}(s) = \log t_n^{\text{old}}(s) + \log q^{\text{new}}(s) - \log q^{\text{old}}(s), \qquad t_n^{\text{new}}(s) \leftarrow \frac{\exp(\log t_n^{\text{new}}(s))}{\sum_{s'} \exp(\log t_n^{\text{new}}(s'))}. \tag{31}$$

The code-level adapters are `cavity(global_posterior, site_factor)` and `update_factor(old_site_factor, old_global_posterior, new_global_posterior)`. The old keywords `posterior`, `factor`, `old_factor`, `old_posterior`, and `new_posterior` are accepted only with a `DeprecationWarning`; mixed canonical/old calls fail closed. Recombination is tested by normalizing $q_{-n}(s)t_n(s)$ and checking recovery of $q(s)$ to floating-point tolerance. A transported site factor is represented as a pmf, so its arbitrary positive natural-parameter scale is fixed by the explicit normalization above.

13.4 Statistical notation and nesting

Symbol	Contractual meaning
n_{seed}	Number of independently seeded worlds/replicates in a declared cell; the inferential unit for seed-level summaries.
n_{trial}	Number of trials nested within one seed and cell; trials are averaged before seed-level inference.
$\Delta = b - a$	Matched robust-minus-naive contrast for the same seed/trial or the declared seed-level reduction.
r_{rb}	Wilcoxon matched-pairs rank-biserial effect, primary standardized effect.
d_{eq}	$2r_{\text{rb}}/\sqrt{1 - r_{\text{rb}}^2}$, a secondary rank-biserial-derived display/planning d-equivalent, not raw Cohen's d.
$\text{CI}_{1-\alpha}$	Percentile bootstrap interval for the named estimand, resampling the declared replication unit.
MCSE	Monte Carlo standard error/precision diagnostic for a simulation summary; it is not a confidence interval.
MDE	Observed-design minimum detectable effect diagnostic under its stated approximation; it is not confirmatory evidence.
p, q	Raw p-value and BH-adjusted q-value; the family and ownership are declared with every report.

For the robustness sweep, the primary result is r_{rb} and the matched mean difference $\overline{\Delta}$ with its bootstrap interval. The d-equivalent is retained only as a monotone secondary display and planning input. When $|r_{\text{rb}}| = 1$, the transform diverges; reports use a finite sentinel and captions disclose saturation rather than presenting a million-scale number as a scientifically interpretable effect. Power, prospective sample size, MCSE, and MDE are observed-effect planning/precision diagnostics, not evidence that a confirmatory effect exists.

The predeclared headline display rule is the largest positive r_{rb} among robust methods, with the declared method order as a deterministic tie break. A report must also expose the complete tied-method set, the tie-break, the method with the largest mean $\overline{\Delta}$, and the method with the largest mean at the worst rate. These are distinct summaries; none is a unique scientific winner when the evidence is tied or conditional.

For the review grid, every configured robust method remains an inferential member and a displayed rate-profile curve. No pooled-mean selection creates a curve, interval, or hypothesis-test member for that surface.

13.5 Code and manuscript naming map

Retired/ambiguous name	Canonical name	Compatibility rule
`beliefs`, `agent_beliefs`	`local_posteriors`	Warned keyword/property adapters; no silent reinterpretation.
`weights`	`base_weights`	Warned keyword adapter. The federation wire key `agent_weights` remains unchanged.
`agent_weights` result property	`normalized_effective_weights`	Warned property adapter; serialized wire compatibility is preserved.
Variational `agent_weights` argument	`raw_effective_weights`	Warned objective API adapter; reports use the canonical term.
`shared_beliefs`	`shared_posteriors`	Warned diagnostics property adapter.
`loss_vec`	`loss_by_state`	Warned generalized-Bayes keyword adapter.
`cohens_d_from_rank_biserial`	`d_equivalent_from_rank_biserial`	Warned function adapter; the returned value is not raw Cohen's d.
Report `cohens_d`	Report `d_equivalent`	New reports are canonical and schema-versioned; readers must fail closed on an unsupported version.

The wire-level key `agent_weights` is preserved because it is a federation transport contract, not a claim about the scale or meaning of the new result fields. A future wire migration requires an explicit version and a fail-closed reader; it must not silently reinterpret the key.

13.6 Source and evidence boundaries

The Friston belief-sharing equations are source equations/protocol claims. Only under the explicit finite-shared-support, posterior-log-potential, and fixed-weight bridge above does the categorical log-linear pool specialize the source message-combination term; the tested $c = 0$ identity remains project-local. The generalized-Bayes and loss limits are implementation analogues checked in the finite categorical model. The contamination, gallery, onset, conditional-world, and review-grid quantities are conditional simulation evidence over declared cells. None of these finite surfaces is an external-data replication, a reconstructed source protocol, a universal attack taxonomy, a causal intervention, or a proof of a server-side robustness

guarantee. Open theory, calibration, protocol, continuous-state, external-data, authenticated-federation, and clean-release work remains open in the project TODO and claim-audit documents.

13.7 Moving sentinel world: communication benefit depends on field of view

The hidden-state/action relation for this extension is summarized in the categorical loop schematic fig. 5; the results below remain the executed moving-world comparisons, not a claim that the schematic's full loop is present in every flat belief-sharing study.

The static sentinel world lets every agent observe the same shared latent, so belief sharing is a refinement rather than a requirement. To stress the *necessity* of communication we add movement and **disjoint** fields of view. The world is a linear grid of 4 cells holding a single binary threat — left half (state 0) or right half (state 1). The 2 sentinels start at evenly tiled positions and each observe a half-open window of cells, so in the default setup agent 0 watches the left half and agent 1 the right half: their views do not overlap. Each agent's likelihood is a confident, signed presence reading for the half it can see, and three control paths (stay / left / right) let it reposition. The expected-free-energy policy scores each candidate move by the expected posterior entropy after one observation and takes the most information-seeking step.

We run 960 trials of 6 steps each under three conditions: *isolated* (random moves, no sharing), *communicating* (random moves plus a log-linear-pool consensus each step), and *EFE-guided* (information-seeking moves plus the same sharing). The measured consensus accuracies are 0.999 (isolated), 0.977 (communicating), and 0.978 (EFE-guided), with a communicating free-energy gap of -0.528 nats relative to the isolated baseline (negative: no free-energy advantage over isolated in this binary-complement regime of logically complete half-views) (fig. 23).

Across 128 independent seeds the EFE-guided accuracy is 0.983 (95 % CI 0.982–0.984), the communicating (random-moves + sharing) accuracy is 0.982 (95 % CI 0.982–0.983), and the isolated accuracy is 0.999 (95 % CI 0.999–0.999). In this binary-complement regime the isolated condition is in fact significantly *higher* on accuracy than the EFE-guided sharing condition — their 95 % intervals do not overlap — and the EFE-vs-isolated accuracy contrast yields Wilcoxon signed-rank $p = 0.0000$ (significant; isolated higher), effect size $r = 1.000$ (large). Sharing is therefore not merely unnecessary in this regime; it costs a small but reliable amount of accuracy. Nor does sharing lower free energy here: the EFE free-energy gap (isolated surprise minus the EFE-guided condition's surprise) is -0.368 nats (95 % CI -0.384—0.353) — negative, so the pooled consensus is slightly *more* surprised by the true state than the isolated baseline, because a single agent's view already suffices. The accuracy case for *necessity* is therefore made only in the larger-state-space disjoint-FOV extension below, not by these binary-complement numbers.

We report these numbers as measured, not assumed. The binary world carries a logical complement: ruling out one's own half implies the other, so a single agent's "not detected" still carries information about the global state, and an isolated agent is not strictly blind. By design, the intended *cannot-decide-alone* regime is the high-noise, few-step corner where one sensor's evidence cannot overcome the flat prior; there belief sharing is meant to fuse the two complementary views into a decisive consensus. That regime is not separately measured here — the construction, the three actions, the EFE rule, and the exact condition protocol are detailed in the supplement (sec. 13.8).

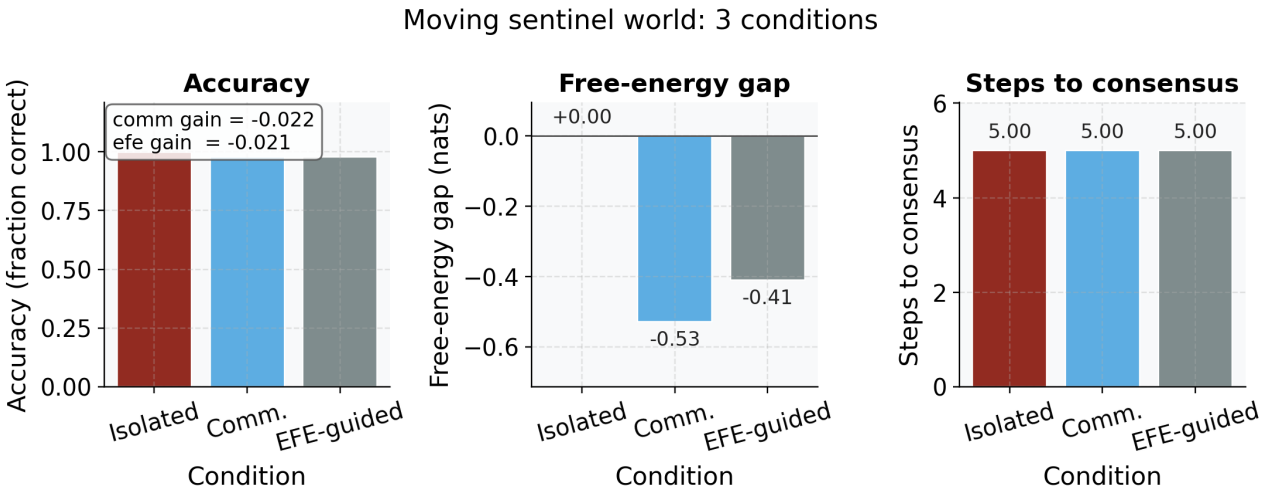


Figure 23: Source relation: original project schematic for the moving-world protocol; estimand: condition-level consensus accuracy, signed free-energy gap, and steps-to-consensus proxy in the stated native units; uncertainty: deterministic seeded run, so no resampling interval is shown. Moving sentinel world across the three conditions (x-axis is condition: isolated, communicating, EFE-guided). Left panel: y-axis shows consensus accuracy (fraction of 960 trials whose pooled argmax matches the truth). Center panel: y-axis shows the signed free-energy gap in nats (isolated surprise minus the condition's surprise on the true state, so a positive value would mean lower free energy than isolated; the measured gaps are negative, plotted against a zero reference line and annotated per bar). Right panel: y-axis shows a coarse steps-to-consensus proxy, with per-bar value annotations showing the three conditions are essentially tied. Each colony runs 6 steps over a 4-cell linear grid with 2 disjoint-FOV agents. Deterministic seeded run, so the bars carry no error band.

13.7.1 Disjoint field-of-view extension

To test whether communication is necessary (not merely beneficial) when observations are non-overlapping, we extend Study 5 to 3 agents each observing a 2-position disjoint window of a 6-position state space (chance-level accuracy 0.167). Isolated agents achieve mean accuracy 0.35 — above chance but far from decisive, since no single agent can infer the global state from a partial window alone. Communicating agents pool complementary beliefs to reach 0.55 (gap 0.19).

This is now a powered result, not an illustrative point estimate. Across 128 independent seeds the isolated accuracy is 0.326 (95% CI 0.320–0.332) and the communicating accuracy is 0.493 (95% CI 0.487–0.499), both clearing the 0.167 chance baseline — isolated agents are *not* at chance, since a partial FOV plus majority voting still carries some signal. The paired Wilcoxon signed-rank test (communicating vs. isolated, matched by seed) gives $p = 0.0000$, effect size $r = 1.000$ (large): communicating beats isolated on every one of the 128 seeds, which is also why the p-value is at the smallest a 128-seed paired sign test can report — it should be read as "every seed agreed," not as a precise magnitude of evidence beyond that floor. Given isolated performance is above chance, the precise claim is not that communication is *logically* necessary for any signal at all, but that it is necessary to approach the communicating-level accuracy under fully disjoint observations: the gap between the two conditions is significant, large, and reproducible, unlike the binary-complement contrast above.

We separately quantify EFE-guided navigation rather than asserting an unquantified "widens the gap" effect. In a matched but smaller-scale disjoint-FOV movement-policy comparison (2 agents, 4-position binary-state grid, belief sharing active in both arms), EFE-guided accuracy is 0.975 versus 0.977 for random movement ($p = 0.1046$, negligible effect): the two movement policies are not significantly different, because both are already near ceiling once belief sharing is active. We report this as the null result it is rather than claiming an unmeasured EFE benefit. fig. 24 summarizes the necessity result.

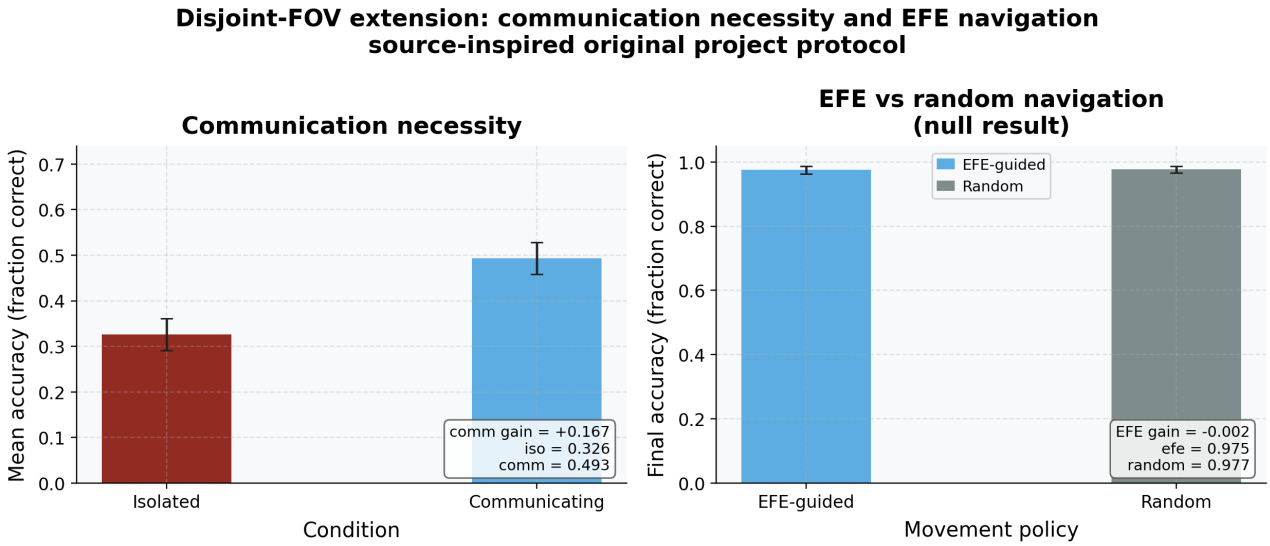


Figure 24: Source relation: source-inspired original project extension of the moving-world mechanism; estimand: condition-level consensus accuracy in the two declared disjoint-FOV protocols; uncertainty: across-seed standard-deviation error bars. Disjoint-FOV extension of the moving sentinel world, as a two-panel figure whose panels come from two separately configured experiments. Left panel (communication necessity, 3 agents each observing a 2-position non-overlapping window of the 6-position world): the x-axis is the condition (isolated vs. communicating); the y-axis is consensus accuracy — drawn as accuracy, the fraction of trials whose pooled argmax matches the true state. The accuracy gap between communicating and isolated conditions quantifies the necessity of belief sharing under fully disjoint fields of view, now backed by the paired Wilcoxon test in the text ($p = 0.0000$) rather than a single point estimate. Right panel (EFE vs random navigation, a smaller 2-agent, 4-position configuration): the x-axis is the movement policy (EFE-guided vs. random); the y-axis is final consensus accuracy — the panel is titled as a null result because that is what it shows. Both policies sit near ceiling once belief sharing is active, so the EFE-guided vs. random contrast is the null result reported in the text ($p = 0.1046$). In both panels, bars show accuracy averaged across seeds; error bars show the across-seed standard deviation.

13.8 Supplement: moving-world methods and condition definitions

This supplement supplies the construction that sec. 13.7 defers here: exactly how the moving-sentinel world is built, how the three actions and the expected-free-energy policy move an agent, and how the *isolated*, *communicating*, and *EFE-guided* conditions differ. It answers the mechanical question left open in the main section — what precisely is held fixed and what varies across the three conditions whose accuracies are contrasted there — so the reported binary-complement numbers can be read against their generative model rather than taken on trust.

The moving-world generative model is built by `build_moving_world`. A linear grid of 4 cells holds one binary hidden state — the half of the grid (left = state 0, right = state 1) that contains the threat. The 2 sentinels start at evenly tiled positions ($i \cdot \lfloor n_{\text{positions}}/n_{\text{agents}} \rfloor$) and each observe a half-open field-of-view window. With the default setup the two FOVs are disjoint, one per half. Each agent's likelihood is a 2×2 matrix over outcomes (detected / not_detected) given the binary state, with a confident signed reading for the half the agent watches; the transition tensor encodes three deterministic control paths — stay, left (reflecting at cell 0), and right (reflecting at the last cell). The hidden-state prior is uniform.

Action selection has two regimes. The random conditions draw each agent's move uniformly from the three controls. The EFE-guided condition uses `efe_policy_select`: for every candidate move it lands the agent at the deterministic next position, reconstructs the likelihood from that viewpoint, and scores the move by the expected posterior entropy after one observation, $H = \sum_o P(o)\, H(P(s \mid o))$ — taking the entropy-minimizing (information-seeking) step.

We compare three conditions — *isolated* (random moves, no sharing), *communicating* (random moves plus a per-step log-linear-pool consensus), and *EFE-guided* (information-seeking moves plus the same per-step sharing) — over 960 trials of 6 steps each, scoring the pooled consensus against the true state. All numerics are deterministic given the run seed.

13.9 Hierarchical POMDP: federated belief sharing across levels

The flat sentinel world couples all agents at a single latent level — the creature's location. A natural extension is a **2-level hierarchical POMDP** in which location inference (Level 1, L1; 9 states) is coupled to a global *context* variable (Level 2, L2; 2 states: `quiet` / `alert`) that modulates the L1 prior. In the `alert` context the creature is expected near the den (center cell); in the `quiet` context the prior is uniform. Each sentinel runs alternating L1/L2 minimization (`fedference.pomdp.hierarchical_infer`) to infer both its location belief and the current context belief, then the colony federates both levels via a log-linear pool.

We compare two conditions over 960 seeded trials with 4 agents at sensor acuity 0.85:

- **Flat** — agents ignore the hierarchy and infer location under a uniform prior;
- **Hierarchical** — agents run 4 alternating-minimization iterations to couple L1 and L2 beliefs before federating.

The measured location accuracies are 0.982 (flat) and 0.969 (hierarchical), a gap of -0.014. Across 128 independent seeds the hierarchical location accuracy is 0.974 (SD 0.005; 95 % CI 0.973–0.975) versus flat 0.982 (95 % CI 0.981–0.982), a mean accuracy gap of -0.008 (95 % CI -0.009—0.007; Wilcoxon signed-rank $p = 0.0000$, effect size $r = 0.940$, large). On location the gap is small but statistically reliable in the *negative* direction — the paired test rejects at $\alpha = 0.05$ and the gap's confidence interval (-0.009—0.007) excludes zero on the negative side — so the hierarchy does not improve location accuracy in this regime; if anything it pays a small, consistent location cost for carrying the extra latent level. Its added value is that it *also* infers the context latent, at accuracy 0.763 against a two-state chance baseline of 0.5. Two-level federation therefore runs L1/L2 inference end-to-end and resolves context above chance while paying a small, reliable location cost relative to the flat baseline (fig. 25). Context beliefs across the alternating-minimization iterations are shown in the top-middle panel: P(alert) sits above the two-state chance line and is stable from the first iteration onward when the observed location is the center cell — the center-cell observation pins the context posterior immediately, because the alert context-conditioned L1 prior is peaked there. The full construction and parameter sweep are detailed in the supplement (sec. 13.10). For the effect of acuity and colony size on these results, see sec. 13.13.

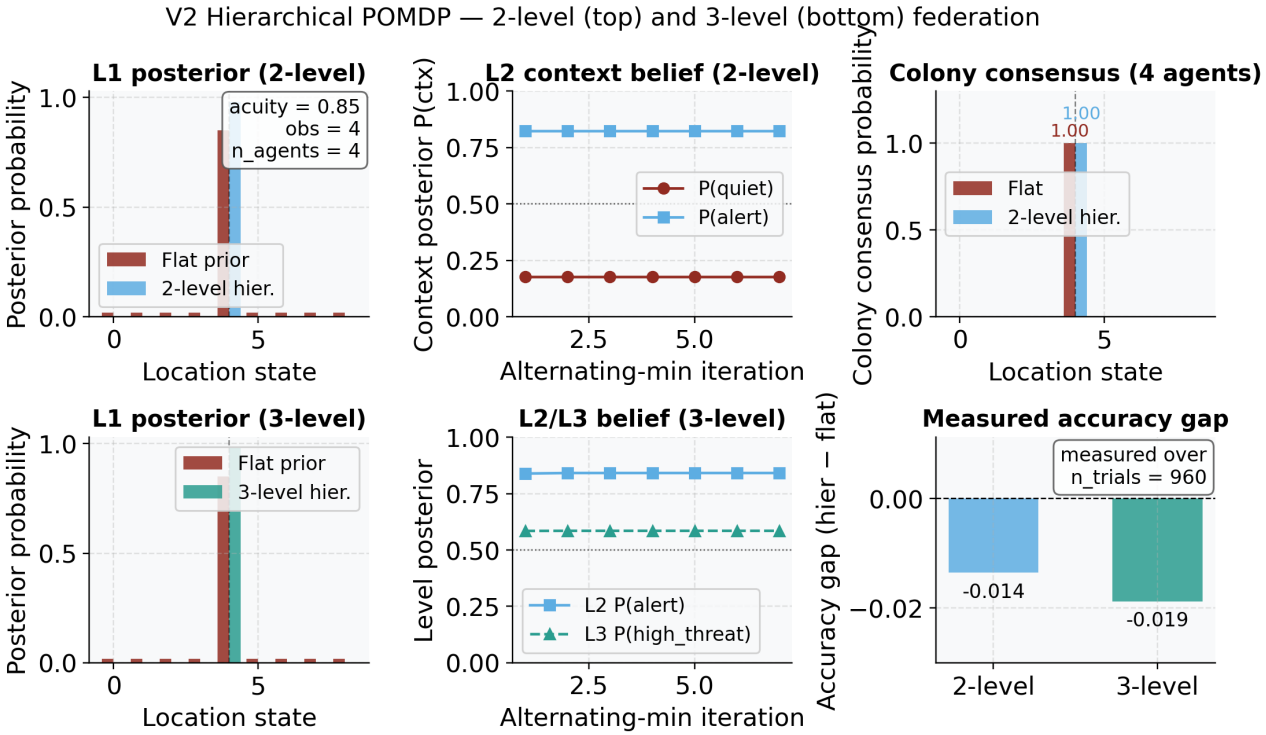


Figure 25: Source relation: source-inspired original project diagnostic for a hierarchical POMDP extension; estimand: posterior probabilities and final location-accuracy gap in the declared seeded protocol; uncertainty: deterministic seeded run, so no resampling interval is shown. Six-panel (2x3) visualization of the V2 hierarchical POMDP belief dynamics. Top row shows the 2-level world; bottom row shows the 3-level extension. Top-left panel: x-axis indexes the 9 location states; y-axis shows posterior probability for the flat-prior and 2-level hierarchical conditions given a single center-cell observation. Top-middle panel: x-axis is alternating-minimization iteration number; y-axis shows the L2 context posteriors P(quiet) and P(alert) under 2-level inference, pinned by the center-cell observation and stable from the first iteration onward. Top-right panel: x-axis indexes location states; y-axis shows the colony L1 consensus probability after federating 4 agents, comparing flat vs 2-level hierarchical. Bottom-left panel: x-axis indexes location states; y-axis shows posterior probability for the flat-prior and 3-level hierarchical conditions given a single center-cell observation. Bottom-middle panel: x-axis is alternating-minimization iteration number; y-axis shows L2 P(alert) and L3 P(high_threat) under 3-level inference, stable across iterations. Bottom-right panel: two bars showing the measured final location-accuracy gap (hierarchical minus flat) for the 2-level and 3-level systems, each a single scalar measured over 960 trials, with a zero reference line. Deterministic seeded run (seed 0), so bars carry no error band.

13.10 Supplement: hierarchical POMDP methods and parameters

This supplement makes the two-level construction of sec. 13.9 concrete: how location (L1) is coupled to context (L2) through context-conditioned priors, what the alternating-minimization update actually computes, and the exact parameters of the executed run. It answers the question the main section brackets — *by what mechanism does a second latent level enter the inference at all* — and thereby fixes why the hierarchy resolves context above chance while leaving location accuracy statistically unchanged.

13.10.1 Generative model for context-gated location inference

The two-level POMDP implemented in `fedference.pomdp.build_hierarchical_world` couples the sentinel's 9-location L1 factor to a 2-state L2 context factor via **context-conditioned L1 priors**:

- **L1 (location)** — the standard 3x3 grid of `build_sentinel_world` with `n_s = 9` states and sensor acuity 0.85;
- **L2 (context)** — a binary state (`quiet` / `alert`) with a symmetric transition matrix (persistence 0.90) and an initial uniform prior;
- **L1 priors given context** — `quiet`: uniform over all 9 cells; `alert`: mass 0.60 at the center cell (the den), the residual spread uniformly.

13.10.2 Inference algorithm for top-down empirical priors

`fedference.pomdp.hierarchical_infer` performs 4 passes of alternating minimization:

1. **L2 → L1 empirical prior**: $\tilde{\pi}_{0,\mathrm{L1}} = \sum_c q_{\mathrm{ctx}}[c]\,\pi_{0,\mathrm{L1}|c}$ (a soft mixture of the two context-conditioned priors).
2. **L1 update**: one-step variational posterior $q_{\mathrm{loc}} = \mathrm{softmax}(\log \tilde{\pi}_{0,\mathrm{L1}} + \log A[\mathrm{obs}, \cdot])$.
3. **L1 → L2 marginal evidence**: $\ell_c = \log(\pi_{0,\mathrm{L1}|c}^{\top} A[\mathrm{obs}, \cdot])$ (evidence for context c from the observed location likelihood).
4. **L2 update**: $q_{\mathrm{ctx}} = \mathrm{softmax}(\log \pi_{0,\mathrm{L2}} + \ell)$.

After 4 iterations the agent broadcasts both q_{loc} and q_{ctx}; the colony federates each level independently via a log-linear pool (eq. 6).

13.10.3 Study parameters for the hierarchical condition

Table 13: Study 6 hierarchical POMDP execution parameters: agent count, seeded trial budget, observation acuity, alternating-minimization iterations, and the L2/L1 state cardinalities used by the two-level condition.

Parameter	Value
Agents	4
Trials	960
Acuity	0.85
Alternating-min iterations	4
L2 context states	2
L1 location states	9
Seed	0

The executed hierarchical configuration is summarized in tbl. 13.

13.11 Three-level hierarchical POMDP: an executed test of the N-level template

The 2-level hierarchical POMDP (sec. 13.9) couples location inference to a single global context. The N-level architecture (`fedference.pomdp.build_nlevel_world`) provides a parameterized stack of levels; the canonical 3-level example couples location (L1; 9 states) to a context variable (L2; 2 states: `quiet` / `alert`) and further to a meta-context variable (L3; 2 states: `low_threat` / `high_threat`) that gates the L2 prior.

$$\tilde{D}_{\mathrm{L2}} = \sum_k q_{\mathrm{L3}}[k]\, p_{\mathrm{L2}|\mathrm{L3}}[k] \tag{32}$$

$$\tilde{D}_{\mathrm{L1}} = \sum_c q_{\mathrm{L2}}[c]\, p_{\mathrm{L1}|\mathrm{L2}}[c] \tag{33}$$

The inference algorithm (`fedference.pomdp.nlevel_infer`) performs 4 passes of top-down / bottom-up alternating minimization: the top-down pass propagates empirical priors from L3 → L2 → L1 via eq. 32 and eq. 33; the bottom-up pass updates each level's belief from the marginal evidence contributed by the level below.

We compare two conditions over 960 seeded trials with 4 agents at sensor acuity 0.85:

- **Flat** — agents ignore all hierarchy and infer location under a uniform prior;
- **3-level** — agents run 4 alternating-minimization iterations across all three levels before federating.

The measured location accuracies are 0.984 (flat) and 0.966 (3-level), a gap of -0.019. Across 128 independent seeds the 3-level location accuracy is 0.976 (SD 0.005; 95 % CI 0.976–0.977) versus flat 0.981 (95 % CI 0.980–0.981), a mean accuracy gap of -0.004 (95 % CI -0.005—0.003; Wilcoxon signed-rank $p = 0.0000$, effect size $r = 0.724$, medium; the location gap over the flat baseline is not statistically significant at this seed count). The 3-level condition additionally reports context accuracy 0.697 and meta-context accuracy 0.547. Against the two-state chance baseline of 0.5, location is recovered and the intermediate context latent is resolved well above chance, but the meta-context latent is only marginally above chance — the weakest

of the three levels — and is therefore *not* convincingly recovered here. The study thus demonstrates that the generic N-level alternating-minimization runs and federates end-to-end and recovers the fastest (location) and intermediate (context) latents; full recovery of the slowest (meta-context) level is left open. The full figure comparing 2-level and 3-level belief dynamics is fig. 25. The declarative layer specification used by the generic constructor is documented in the supplement (sec. 13.12). For the effect of acuity and colony size on these results, see sec. 13.13.

13.12 Supplement: N-level hierarchical POMDP methods

This supplement specifies the generic N-level architecture that sec. 13.11 exercises at depth three: how the meta-context (L3), context (L2), and location (L1) factors are chained through conditioned priors, what the declarative `LayerSpec` interface fixes versus leaves free, and the top-down/bottom-up passes the inference runs. It answers *what the executed 3-level result is a special case of* — the reason the same code runs at other depths without new mathematics — while recording that only the declared 3-level configuration is empirically evaluated here.

13.12.1 Generative model for an N-level hierarchy

The 3-level POMDP implemented in `fedference.pomdp.build_3level_world` extends the 2-level construction (sec. 13.10) by adding a top-level meta-context factor:

- **L3 (meta-context)** — 2 states (`low_threat` / `high_threat`) with initial uniform prior, gating the L2 context prior;
- **L2 (context)** — 2 states (`quiet` / `alert`) with context-conditioned L1 location priors, gating the L1 prior;
- **L1 (location)** — the standard 3x3 grid with 9 states and sensor acuity 0.85.

The conditioned priors are (see eq. 32 and eq. 33):

L3 state	L2 prior (quiet, alert)
`low_threat`	(0.50, 0.50) — uniform context
`high_threat`	(0.20, 0.80) — peaked at alert

L2 state	L1 prior
`quiet`	uniform over all 9 location states
`alert`	mass 0.60 at center cell (flat index 4), residual uniform

13.12.2 Generic N-level architecture

`fedference.pomdp.LayerSpec` and `fedference.pomdp.build_nlevel_world` implement the generic N-level version. The declarative layer specification is stored at `src/fedference/config/hierarchical_layers.yaml` and mirrors the canonical 3-level defaults (a standalone documentation artifact not read by any code path, kept in sync with the `build_3level_world` defaults). The constructor accepts depth >= 2; the executed empirical result in this manuscript is restricted to the declared 3-level configuration, and the leaf layer must carry `n_states == N_LOCATIONS`.

13.12.3 Inference algorithm across hierarchy levels

`fedference.pomdp.nlevel_infer` performs 4 passes of top-down / bottom-up alternating minimization over all N levels:

1. **Top-down pass** — compute the empirical prior for each level by marginalizing over the level above (eq. 32, eq. 33).
2. **L1 update** — one-step variational posterior on the observation: $q_{\text{loc}} = \text{softmax}(\log \tilde{\pi}_{0,\text{L1}} + \log A[\text{obs}, \cdot])$.
3. **Bottom-up pass** — update each non-leaf level's belief from the marginal evidence contributed by the level below: $\ell_j = \log(\tilde{p}^{\top}_{\text{child}|\text{parent}=j} q_{\text{child}})$.

After 4 iterations the agent broadcasts all N level beliefs; the colony federates each level independently via a log-linear pool (eq. 6).

13.12.4 Study parameters for the three-level run

Table 14: Study 7 three-level hierarchical POMDP execution parameters: agent count, seeded trial budget, observation acuity, alternating-minimization iterations, and the L3/L2/L1 state cardinalities used by the three-level condition.

Parameter	Value
Agents	4
Trials	960
Acuity	0.85
Alternating-min iterations	4
L3 meta-context states	2
L2 context states	2
L1 location states	9
Seed	0

The executed three-level configuration is summarized in tbl. 14.

13.13 Parameter sensitivity of federation benefit

Studies 1–7 fix specific parameter configurations (sensor acuity, colony size) to isolate mechanistic claims. A natural question is whether the federation benefit is **robust to those choices** or is an artifact of a narrow operating point. Study 8 addresses this with a systematic 2-D sensitivity sweep over sensor acuity and colony size.

We sweep sensor acuity $\kappa \in \{0.40, 0.55, 0.70, 0.85, 0.95\}$ and colony size $n \in \{2, 4, 6, 8, 10\}$, evaluating two systems:

- **Belief sharing** (Study 1 architecture) — accuracy gap = communicating minus isolated mean accuracy;
- **Hierarchical POMDP** (Study 6 architecture) — accuracy gap = hierarchical minus flat location accuracy.

Each cell averages $n_{\text{trials}} = 20$ independent trials to reduce Monte-Carlo noise at the cell level.

The resulting heatmaps (fig. 26) show the accuracy gap for both systems as a function of acuity and colony size. Green cells indicate that federation benefits the colony; red cells indicate that the chosen configuration yields no benefit or a slight deficit. The symmetric RdYlGn colormap is centered on zero so the sign of the benefit is immediately legible.

Across the grid the following patterns hold:

1. **The belief-sharing benefit lives at low-to-moderate acuity.** The accuracy gap peaks in the second-lowest acuity row — where individual observations carry some signal but no single sentinel resolves the location alone, so pooled evidence pays most — and remains uniformly positive across the two lowest-acuity rows for colonies of at least four agents. Near ceiling acuity the gap shrinks toward zero: single agents already solve the task, so federation has nothing left to add.
2. **Colony size acts through a floor, not a smooth slope.** The two-agent column shows an exactly zero belief-sharing gap at every acuity: under self-exclusion ("agents do not hear themselves"), each member of a two-agent colony hears exactly one incoming belief, so the heard consensus adds no pooled evidence. Colonies of four or more realize the low-acuity benefit.
3. **The hierarchical gap is near zero across most of the grid.** The hierarchical-minus-flat location-accuracy gap is approximately zero over most cells, with a few strongly negative low-acuity cells and small positive cells confined to the two-agent column — consistent with the per-study finding that the hierarchical architecture matches rather than beats the flat baseline on location accuracy.

The full parameter grid and protocol details are in the supplement (sec. 13.14). A native-unit cross-study overview of the headline metrics across all 9 studies is shown in fig. 27.

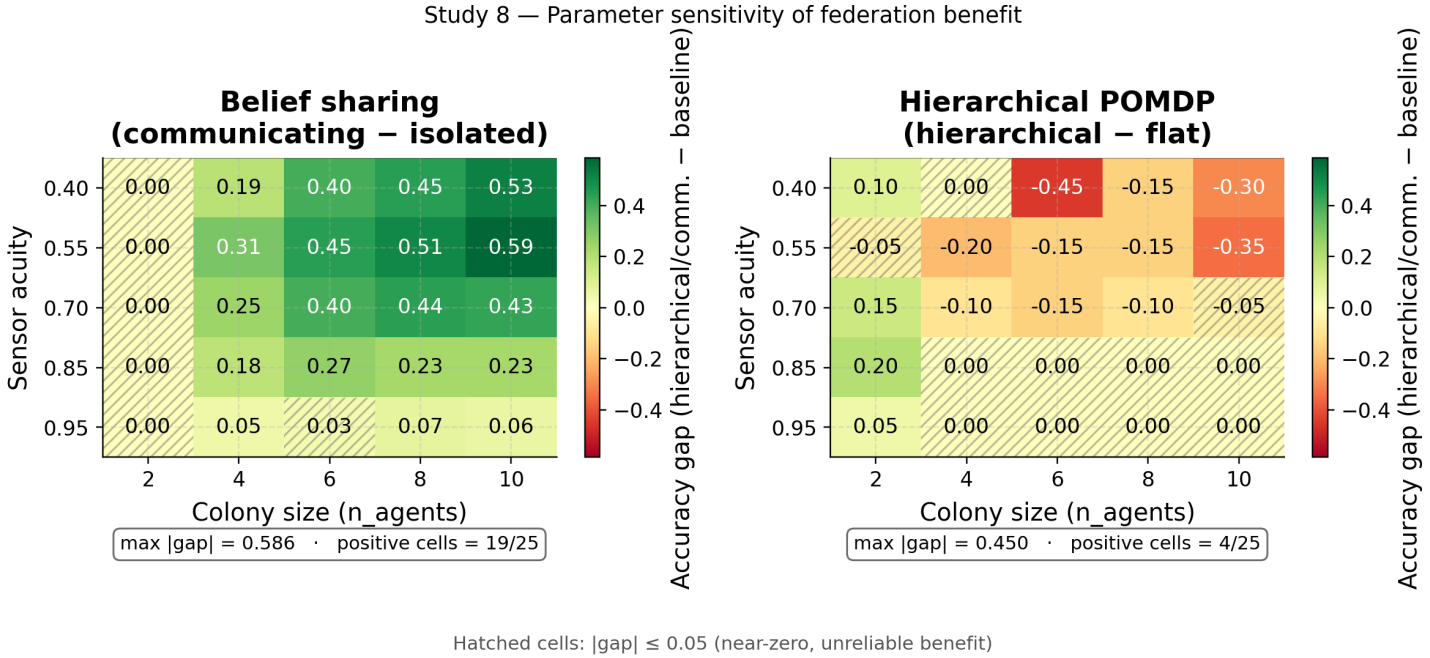


Figure 26: Source relation: original project sensitivity diagnostic; estimand: per-cell accuracy gaps (fractions) as functions of acuity and colony size; uncertainty: deterministic per-cell means over the declared trials, with no resampling interval. Two-panel heatmap (1×2) of the Study 8 parameter sensitivity sweep. Left panel: y-axis indexes sensor acuity (0.40–0.95, 5 levels); x-axis indexes colony size (2–10 agents, 5 levels); color encodes the belief-sharing accuracy gap (communicating minus isolated mean accuracy). Right panel: identical axes; color encodes the hierarchical POMDP location accuracy gap (hierarchical minus flat). Color scale: RdYlGn symmetric around zero — green denotes federation benefit, red denotes deficit. Diagonal hatching marks cells with $|\text{gap}| \leq 0.05$ (unreliable, near-zero benefit). Cell values are deterministic per-cell means over 20 trials; no resampling error band is shown — the sweep protocol is detailed in the sensitivity supplement.

13.14 Supplement: parameter-sensitivity methods

This supplement documents how the sensitivity grid of sec. 13.13 is generated and — importantly for a reader trying to reconcile numbers across studies — where its two sweeps use *different* seeding and trial budgets. It answers two questions the main section leaves implicit: exactly which seed drives each cell (so any cell is independently reproducible), and why the cross-study summary's sensitivity row is not directly comparable, at matched trial counts, to the standalone heatmap.

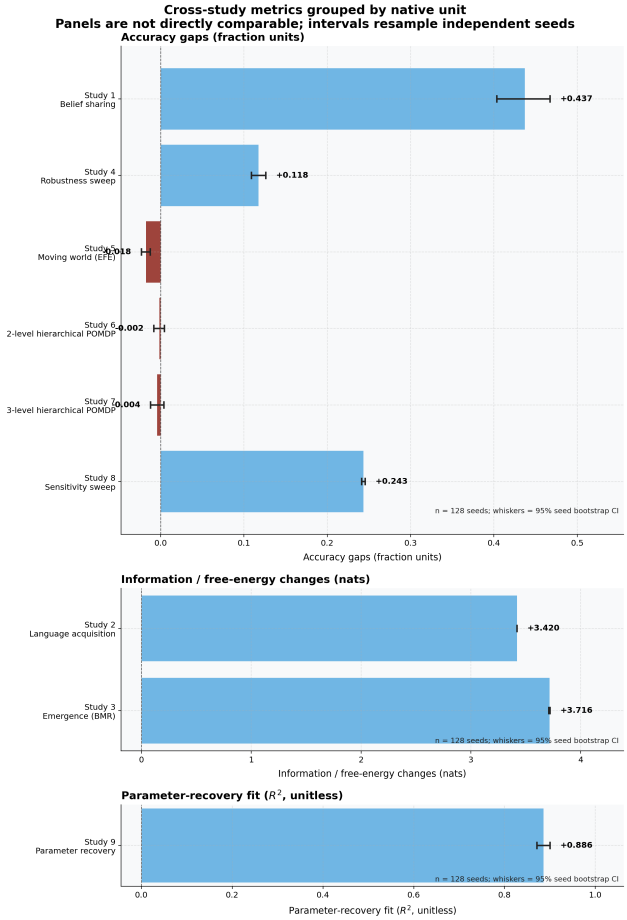


Figure 27: Source relation: original project cross-study summary; estimand: grouped study-level means in native units (accuracy fractions, nats, or R^2), never a cross-unit ranking; uncertainty: seed-level bootstrap confidence intervals. Horizontal native-unit facet chart summarizing the key federation benefit metric for each of the 9 studies (Studies 1–9). x-axis indexes benefit value in metric-specific units (accuracy gain for Studies 1, 4, 5, 6, 7, 8; KL reduction for Study 2; ΔF for Study 3; R^2 for Study 9). y-axis lists the 9 studies (one row each), ordered from Study 1 at the top to Study 9 at the bottom within each native-unit facet. Each mark shows the mean over 128 independent seeds with intervals spanning the 95 % bootstrap confidence interval. There is no cross-unit ranking: zero is a within-unit reference only, and the facet labels carry the units. Consistent with the per-study results, Studies 5 (moving world, EFE) and 6 (2-level hierarchical) sit at approximately zero within their respective units.

13.14.1 Experimental protocol for grid sensitivity

The sensitivity sweep is implemented in `fedference.experiments.run_belief_sharing_sensitivity` and `fedference.experiments.run_hierarchical_sensitivity`. Each function accepts a tuple of acuity values and a tuple of colony sizes. In the **belief-sharing** sweep every (acuity, colony-size) cell averages n_{trials} independent trials, each seeded via a deterministic formula:

$$\text{seed}_{\text{cell}} = \text{seed}_{\text{base}} + i \cdot 10^5 + j \cdot 10^3 + t \tag{34}$$

The deterministic seed rule eq. 34 makes every grid cell and trial independently reproducible from the base seed.

where i indexes acuity, j indexes colony size, and t indexes the trial within a cell. For the belief-sharing sweep this guarantees that:

1. no two cells share a trial seed (no correlation between cells);
2. re-running with the same `seed_base` is bit-identical (reproducibility);
3. different `seed_base` values produce independent replicates (robustness checking).

The **hierarchical** sweep uses a simpler protocol: every cell calls `run_hierarchical_world` once with the same base seed (its internal trials are seeded by that run), so hierarchical cells share the base seed rather than the per-cell formula above.

13.14.2 Grid parameters for acuity and colony size

Parameter	Values
Sensor acuity κ	{0.40, 0.55, 0.70, 0.85, 0.95}
Colony size n	{2, 4, 6, 8, 10}
Trials per cell	20
Base seed	0

The 5×5 = 25 cells per system are run with `seed_base = 0` by default; `generate_sensitivity_heatmap` accepts a `seed` argument to override this.

13.14.3 Belief-sharing condition in the sensitivity grid

Each trial in the belief-sharing sweep:

1. Draws a random true state and one noisy observation per agent (same protocol as `run_belief_sharing`).
2. Runs one belief-sharing round with `communicate=True` (communicating) and `communicate=False` (isolated).
3. Records `mean_accuracy` for each condition.
4. The cell value is the average over `n_trials` of this gap.

13.14.4 Hierarchical POMDP condition in the sensitivity grid

Each cell in the hierarchical sweep calls `run_hierarchical_world` once with the cell's acuity and colony size, passing the constant base seed (not the per-cell formula, which applies only to the belief-sharing sweep). The returned `location_accuracy_gap` (hierarchical minus flat) becomes the cell value.

13.14.5 Figure rendering for sensitivity summaries

`generate_sensitivity_heatmap` assembles the two grids into a 1×2 matplotlib `imshow` figure with RdYlGn colormap, symmetric bounds at $\pm \max(|\text{gap}|)$, per-cell numeric annotations, and a per-panel colorbar labeled "Accuracy gap (hierarchical/comm. − baseline)". The figure is written to `../figures/sensitivity_heatmap.png`.

13.14.6 Cross-study summary construction

`generate_cross_study_summary` runs a 128-seed ($n_{\text{seeds}} = 128$) ensemble over Studies 1–9 and reports the mean ± 95 % bootstrap CI of the key federation-benefit metric for each study. The metric definitions are:

The robustness row uses 40 matched trials per seed and rate; the trial-level observations are reduced within seed before the cross-study summary is formed. This preserves the seed as the independent Monte Carlo unit.

The Study 8 row below uses 3 trials per cell — smaller than the full-resolution 20-trial `Trials per cell` grid documented above for the standalone sensitivity heatmap figure, a deliberate runtime budget for the per-seed cross-study loop rather than an oversight — so the two are not directly comparable at matched trial counts.

Study	Metric
1 — Belief sharing	Accuracy gain: communicating − isolated
2 — Language acquisition	KL reduction: initial − final
3 — Emergence (BMR)	ΔF for redundant pruning
4 — Robustness sweep	Accuracy gain: pooled display robust method − naive at worst contamination rate

Study	Metric
5 — Moving world (EFE)	Accuracy gain: EFE-guided − isolated
6 — Hierarchical POMDP (2-level)	Location accuracy gap: hierarchical − flat
7 — 3-level POMDP	Location accuracy gap: 3-level − flat
8 — Parameter sensitivity	Mean accuracy gap across the sensitivity grid
9 — Parameter recovery	R^2 for acuity identifiability

Bootstrap CIs use 5000 resamples (default `n_boot` in `fedference.statistics.bootstrap_ci`).

14 Parameter recovery: acuity selection on the tested grid

Parameter recovery probes whether the executed observation model contains enough information to distinguish sensor acuity under the study design. We sweep acuity values 0.60, 0.70, 0.80, 0.90: for each true acuity the model generates 200 synthetic observations per trial across 960 independent trials, fits acuity by marginal-likelihood grid search, and compares the recovered value with ground truth.

Across the sweep the mean absolute recovery error is 0.0232 and the coefficient of determination of mean-recovered versus true acuity is $R^2 = 0.9999$. Within this finite grid and observation budget, recovered acuity tracks the identity line with the reported error (fig. 28). This is evidence of practical acuity recoverability for the executed design, not a proof of global or structural identifiability and not an acuity-by-colony-size study.

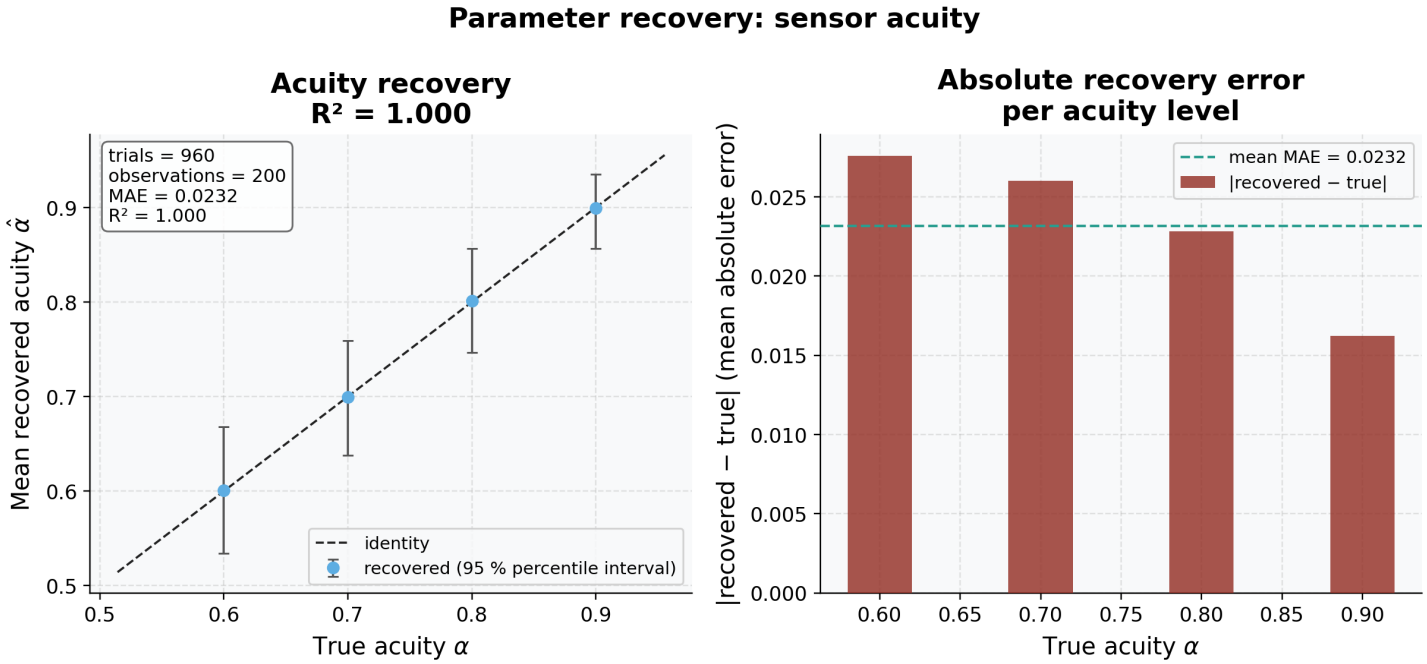


Figure 28: Two-panel parameter-recovery figure. Source relation: original project parameter-recovery diagnostic; estimand: recovered acuity and absolute error in probability units; uncertainty: empirical percentile intervals across independent trials. In the left panel, the x-axis is true acuity and the y-axis is recovered acuity, both in probability units; error bars show the 95% empirical percentile interval across 960 trials per condition, and the diagonal is the identity reference. This interval is a descriptive quantile of the independent-trial estimates, not a bootstrap confidence interval or Bayesian credible interval. In the right panel, the x-axis is tested true acuity and the y-axis is absolute acuity error in probability units; the horizontal line is the global mean absolute error. These finite-grid results quantify acuity recovery for 200 observations per trial; they do not establish global structural identifiability.

14.1 Structure learning: does the hierarchy earn its depth?

Study 7 shows the 3-level agent runs and federates end to end, but a deeper model is only warranted if the extra level carries information. We close the loop with a structure-learning test: given a trained hierarchy and one leaf observation, does Bayesian model reduction correctly decide whether the top meta-context level should be *kept* or *pruned*?

We reduce at the level granularity (:func:`fedference.bayesian_model_reduction.hierarchical_reduce`). For each non-leaf level we measure its **Bayesian surprise** $\mathrm{KL}(q_i \,\|\, \tilde{p}_i)$ — how far the leaf observation moves that level's belief q_i from its top-down prior $\tilde{p}_i$. A level whose belief the data never move carries no structure and is prunable; an informative level moves and is kept. This is an inference-derived divergence, not a model re-fit, so it cannot manufacture a difference the generative model does not contain.

The test is directional by construction. We build two 3-level worlds that differ *only* in the top level's conditioned priors: a **degenerate** world whose meta-context is non-gating (both meta-context states predict the same context distribution) and an **informative** world whose meta-context sharply distinguishes the two contexts. On the degenerate world the top level earns a Bayesian surprise of 0.000 nats and is flagged prunable (recovers the

two-level structure: Yes); on the informative world the same level earns 0.328 nats and is kept (Yes); fig. 29 shows the per-level surprise for both worlds side by side. Because the two worlds share every other parameter, the opposite verdict is attributable to the meta-context's information alone — the reduction discovers the right depth rather than assuming it.

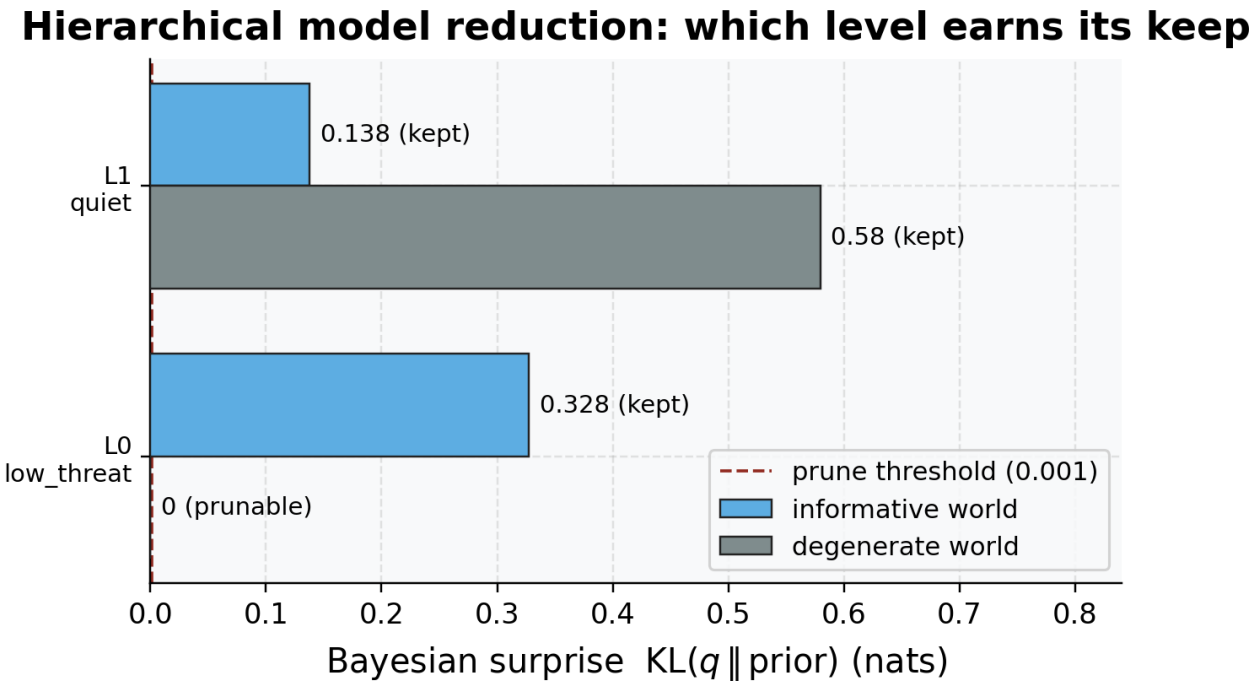


Figure 29: Source relation: original project BMR structure-learning diagnostic related to the mechanism in Friston et al. Fig. 9; estimand: per-level Bayesian surprise in nats and the resulting prune/keep decision; uncertainty: deterministic schematic worlds, so no resampling interval is shown. Per-level Bayesian surprise for the two 3-level worlds. y-axis: the non-leaf reduction targets, indexed top-down from the reduction routine as level 0 = the meta-context (the topmost non-leaf level, L3 in the location-first L1/L2/L3 convention used elsewhere) and level 1 = the context (L2); the leaf location level (L1) is never a reduction target. x-axis: Bayesian surprise $\mathrm{KL}(q \,\|\, \text{prior})$ in nats — the information the leaf observation added at that level. Blue bars: the informative world (top level kept). Grey bars: the degenerate world (top level prunable). The dashed red line is the prune threshold; a level whose surprise falls below it is structurally unnecessary. The degenerate meta-context (grey, level 0) sits at 0.000 nats and is pruned, recovering the two-level model, while the informative meta-context (blue, level 0) at 0.328 nats is retained; both worlds keep the context level (level 1). Deterministic schematic worlds (no resampling), so no error band is applicable.

This is the same Beta-function model-reduction machinery that drives the emergence study (sec. 7.4, eq. 16), lifted from pruning redundant *states* within one level to pruning a redundant *level* of the hierarchy — an honest, tested answer to "how deep should the generative model be?" that the data, not the modeler, decides.

14.2 Sharp server heuristic: influence and finite-breakdown characterization

The server-side `robust_aggregate` rule is the sharp heuristic axis of the three-axes design (sec. 7.2.1). It has BH-rejected positive contrasts in the configured accuracy verdict in sec. 7.5.1 but has declared reversals elsewhere. Unlike the objective-backed `variational_aggregate`, no closed-form free-energy derivation has been established for it in this repository. A separate scoped proposition in the aggregation-objective supplement rules out the declared continuously differentiable, separable forward-KL objective class for the implementation's raw log-pool block; it does not rule out every broader coupled or fixed-point-only construction. The rule therefore remains a heuristic whose positive formal property is bit-identical recovery of the log-linear pool at `robustness = 0` (eq. 7). This section does not promote the scoped negative result into an objective certificate; it *measures* the heuristic empirically, and the measurement makes its honesty boundary concrete.

We measure two things (fig. 30). First, a **numerical influence function**: we drag one agent's belief a growing fraction toward a confident-wrong contamination point and read its converged pooling weight. At `robustness = 0` the weight is a flat $1/n$ at every perturbation — the naive pool never down-weights anyone — which anchors the instrument to the proven recovery corner. At positive robustness the dragged agent's influence falls (not strictly monotonically — a tiny drag can briefly *raise* it before the divergence penalty dominates, an honest non-monotonicity we report rather than smooth away).

Second, and more consequentially, a **breakdown witness**. We add colluding confident-wrong adversaries — all broadcasting the same false state — to a fixed colony of 5 honest sentinels until each aggregator's consensus argmax is *captured* (flips to the adversaries' target). The sharp heuristic is captured by 2 colluders; the conservative objective-backed variational rule withstands more, capitulating only at 4. Both counts are **finite** (Yes): a colluding majority overwhelms either rule. That finite breakdown point is the honest headline: neither rule has an unconditional truth-recovery claim under coordinated collusion. The absence of an objective theorem for `robust_aggregate` is a separate derivational boundary, and the finite capture measurement neither establishes estimator-level B-robustness nor refutes the variational rule's stated raw effective-weight result.

The report also runs a declared diagnostic grid over state dimension, honest-agent count, robustness, four simple attack mechanisms, and balanced versus adversary-downweighted base weights. This is a coverage instrument for finding counterexamples, not a random sample of worlds and not a

theorem search over all simplexes. A finite capture row is evidence against a universal guarantee; an uncaptured row is only "not found within this search budget."

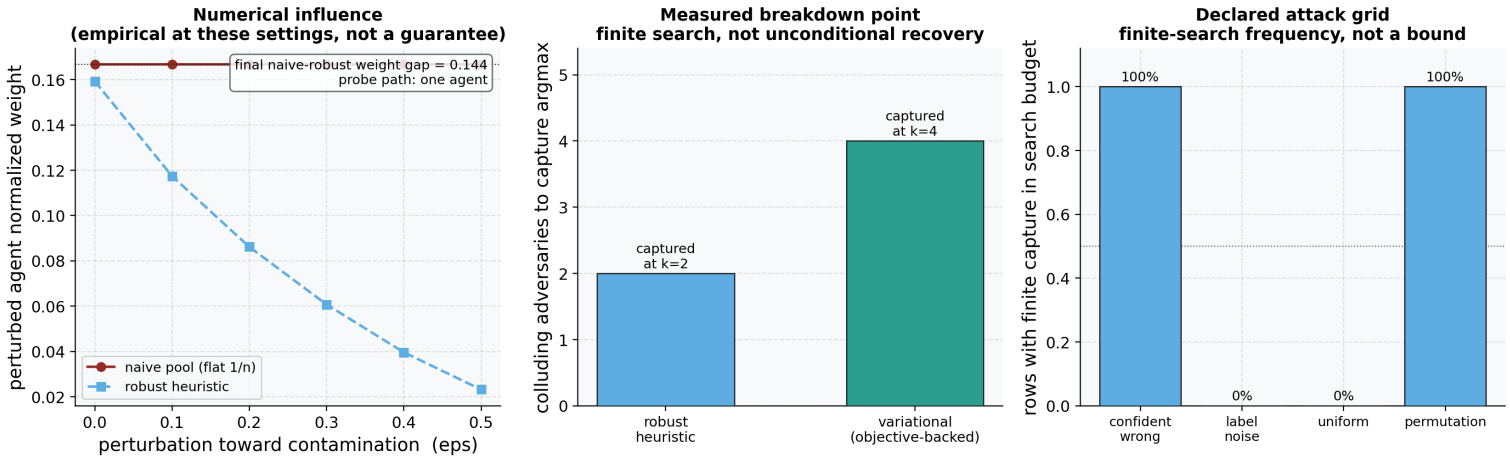


Figure 30: Source relation: original project diagnostic of the server-side heuristic; estimand: numerical influence, finite-search breakdown count, and declared-grid capture fraction; uncertainty: deterministic seeded colonies, so no resampling interval is shown. Empirical characterization of the `robust_aggregate` heuristic (two panels plus an optional attack-grid diagnostic). Left panel (numerical influence): the x-axis is the perturbation fraction by which one agent's belief is dragged toward a confident-wrong contamination point; the y-axis is that agent's converged normalized pooling weight, plotted for the naive pool (flat at $1/n$, dotted reference) and the robust heuristic (down-weighting). The inset reports the final naive-minus-robust weight gap at the end of the probed path. Labeled "empirical, at these settings — not a guarantee." Right panel (measured breakdown point): the x-axis is the aggregator (robust heuristic vs objective-backed variational); the y-axis is the number of colluding confident-wrong adversaries that captures that aggregator's consensus argmax — the robust heuristic at $k = 2$ and the variational rule at $k = 4$. Both bars are finite, so neither rule has an unconditional truth-recovery guarantee against coordinated collusion; this does not negate the variational rule's per-agent effective-weight theorem. Deterministic seeded colonies (no resampling), so no error band is applicable. The optional third panel reports the fraction of declared grid rows with finite capture within the configured adversary budget; it is not a probability or a global breakdown bound.

15 References

The bibliography lives in `manuscript/references.bib` and is read by Pandoc during the PDF render. The build pipeline invokes Pandoc with `--natbib`, so every Pandoc citation marker in the manuscript is rewritten to the appropriate LaTeX citation command and resolved against the bib file. Titles in the bib file are reproduced verbatim, including any British spellings, because they are quotations of the original sources.

To validate that `references.bib` is syntactically clean and contains the required fields per entry type, this validator is only runnable when the project is checked out under the template monorepo's `projects/working/` (it is not on the standalone repo's own dependency graph), invoked from the monorepo root with a monorepo-relative path:

```
uv run python -m infrastructure.reference.citation.cli validate \
    projects/working/active_fedference/manuscript/references.bib --strict
```


Ali E. Abbas. A kullback–leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37, 2009. doi: 10.1287/deca.1080.0133. URL https://doi.org/10.1287/deca.1080.0133.

Mahault Albarracin, Daphne Demekas, Maxwell J. D. Ramstead, and Conor Heins. Epistemic communities under active inference. *Entropy*, 24(4): 476, 2022. doi: 10.3390/e24040476.

Matthew Ashman, Thang D. Bui, Cuong V. Nguyen, Stratis Markou, Adrian Weller, Siddharth Swaroop, and Richard E. Turner. Partitioned variational inference: A framework for probabilistic federated learning, 2022. URL https://arxiv.org/abs/2202.12275.

Dmitry Bagaev, Albert Podusenko, and Bert de Vries. RxInfer: A julia package for reactive real-time Bayesian inference. *Journal of Open Source Software*, 8(84):5161, 2023. doi: 10.21105/joss.05161.

Ayanendranath Basu, Ian R. Harris, Nils L. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998. doi: 10.1093/biomet/85.3.549.

Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. doi: 10.1098/rstl.1763.0053. URL https://doi.org/10.1098/rstl.1763.0053.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Daniel Bernoulli. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5:175–192, 1738. URL https://archive.org/details/SpecimenTheoriaeNovaeDeMensuraSortis.

Jacob Bernoulli. *Ars conjectandi*. Thurneysen, Basel, 1713. URL https://archive.org/details/jacobibernoulli00bern.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f4b9ec30ad9f68f89b29639786cb62ef-Abstract.html.

Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, pages 657–665, 1784. URL https://bibbase.org/network/publication/denbspborda-mmoiresurleslectionsauscrutin-1781.

Thang D. Bui, Cuong V. Nguyen, Siddharth Swaroop, and Richard E. Turner. Partitioned variational inference: A unified framework encompassing federated and continual learning, 2018. URL https://arxiv.org/abs/1811.11206.

Luiz M. Carvalho, Daniel A. M. Villela, Flavio C. Coelho, and Leonardo S. Bastos. Bayesian inference for the weights in logarithmic pooling. *Bayesian Analysis*, 18(1), 2023. doi: 10.1214/22-BA1311.

Jean-Antoine-Nicolas de Caritat Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785. URL https://archive.org/details/essaisurlapplic00conggoog.

Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, 99:102447, 2020. doi: 10.1016/j.jmp.2020.102447.

Abraham de Moivre. *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*. W. Pearson, London, 1718. URL https://archive.org/details/bim_eighteenth-century_the-doctrine-of-chances_moivre-abraham-de_1718.

Franz Dietrich. Fully bayesian aggregation. *Journal of Economic Theory*, 194:105255, 2021. doi: 10.1016/j.jet.2021.105255. URL https://doi.org/10.1016/j.jet.2021.105255.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. doi: 10.1201/9780429246593.

Michael P. Fay and Michael A. Proschan. Wilcoxon–mann–whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, 4:1–39, 2010. doi: 10.1214/09-SS051. URL https://doi.org/10.1214/09-SS051.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787.

Karl Friston and Will Penny. Post hoc bayesian model selection. *NeuroImage*, 56(4):2089–2099, 2011. doi: 10.1016/j.neuroimage.2011.03.062.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: A process theory. *Neural Computation*, 29(1):1–49, 2017. doi: 10.1162/NECO_a_00912. URL https://doi.org/10.1162/NECO_a_00912.

Karl J. Friston, Thomas Parr, Conor Heins, Axel Constant, Daniel Friedman, Takuya Isomura, Chris Fields, Tim Verbelen, Maxwell Ramstead, John Clippinger, and Christopher D. Frith. Federated inference and belief sharing. *Neuroscience & Biobehavioral Reviews*, 156:105500, 2024. doi: 10.1016/j.neubiorev.2023.105500. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11139662/.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008. doi: 10.1016/j.jmva.2008.02.004.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 813–822. PMLR, 2018. URL https://proceedings.mlr.press/v84/futami18a.html.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 1986. doi: 10.1214/ss/1177013825.

Christian Genest, Kevin J. McConway, and Mark J. Schervish. Characterization of externally bayesian pooling operators. *The Annals of Statistics*, 14(2), 1986. doi: 10.1214/aos/1176349934.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2015. doi: 10.1007/s10463-014-0499-0.

Peter Grünwald. The safe bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, pages 169–183. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-34106-9_16.

Conor Heins, Beren Millidge, Daphne Demekas, Brennan Klein, Karl Friston, Iain D. Couzin, and Alexander Tschantz. pymdp: A python library for active inference in discrete state spaces. *Journal of Open Source Software*, 7(73):4098, 2022. doi: 10.21105/joss.04098.

Conor Heins, Beren Millidge, Lancelot Da Costa, Richard P. Mann, Karl J. Friston, and Iain D. Couzin. Collective behavior from surprise minimization. *Proceedings of the National Academy of Sciences*, 121(17):e2320239121, 2024. doi: 10.1073/pnas.2320239121.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2 edition, 2009. ISBN 9780470434697. doi: 10.1002/9780470434697. URL https://doi.org/10.1002/9780470434697.

Christiaan Huygens. *De ratiociniis in ludo aleae*. Frans van Schooten, Leiden, 1657. URL https://archive.org/details/DeRatiociniisInLudoAleae.

Jack Jewson, Jim Q. Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018. doi: 10.3390/e20060442. URL https://doi.org/10.3390/e20060442.

Wenxin Jiang and Martin A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5), 2008. doi: 10.1214/07-AOS547.

Aleksandr Karakulev, Usama Zafar, Salman Toor, and Prashant Singh. Bayesian robust aggregation for federated learning, 2025. URL https://arxiv.org/abs/2505.02490.

Rafael Kaufmann, Pranav Gupta, and Jacob Taylor. Active inference and collective intelligence. In *Cognitive Systems Research*, volume 68, pages 1–13. Elsevier, 2021. doi: 10.1016/j.cogsys.2021.01.001.

B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012. doi: 10.1214/12-EJS675. URL https://doi.org/10.1214/12-EJS675.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022. URL https://jmlr.org/papers/v23/19-1047.html.

Elizabeth Koehler, Elizabeth Brown, and Sebastien J.-P. A. Haneuse. On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162, 2009. doi: 10.1198/tast.2009.0030. URL https://pubmed.ncbi.nlm.nih.gov/22544972/.

Pierre-Simon Laplace. Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences*, pages 621–656, 1774. URL http://sites.mathdoc.fr/cgi-bin/oetoc?id=OE_LAPLACE___8.

Cen-Jhih Li, Pin-Han Huang, Yi-Ting Ma, Hung Hung, and Su-Yun Huang. Robust aggregation for federated learning by minimum γ-divergence estimation. *Entropy*, 24(5):686, 2022. doi: 10.3390/e24050686.

Adam Loy and Jenna Korobova. Bootstrapping clustered data in r using lmeresampler. *arXiv preprint arXiv:2106.06568*, 2021. doi: 10.48550/arXiv.2106.06568. URL https://arxiv.org/abs/2106.06568.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 1273–1282, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Stephen R. Midway. Principles of effective data visualization. *Patterns*, 1(9):100141, 2020. doi: 10.1016/j.patter.2020.100141. URL https://doi.org/10.1016/j.patter.2020.100141.

Terje Mildner, Paris Giampouras, and Theodoros Damoulas. Rates of convergence of generalised variational inference posteriors under prior misspecification. *arXiv preprint arXiv:2510.03109*, 2025a. doi: 10.48550/arXiv.2510.03109. URL https://arxiv.org/abs/2510.03109.

Terje Mildner, Oliver Hamelijnck, Paris Giampouras, and Theodoros Damoulas. Federated generalised variational inference: A robust probabilistic federated learning framework. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 44134–44174. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/mildner25a.html.

Jeffrey W. Miller and David B. Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527): 1113–1125, 2018. doi: 10.1080/01621459.2018.1469995. URL https://doi.org/10.1080/01621459.2018.1469995.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B. Dunson. Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research*, 18(124):1–40, 2017. URL https://jmlr.org/papers/v18/16-655.html.

Pierre Rémond de Montmort. *Essay d'analyse sur les jeux de hazard*. Jacques Quillau, Paris, 1708. URL https://archive.org/details/ldpd_6444894_000.

Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11): 2074–2102, 2019. doi: 10.1002/sim.8086. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6492164/.

Shinichi Nakagawa and Innes C. Cuthill. Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, 82(4):591–605, 2007. doi: 10.1111/j.1469-185X.2007.00027.x. URL https://doi.org/10.1111/j.1469-185X.2007.00027.x.

Hien Duy Nguyen and Jacob Westerhout. Closed-form solutions to some generalized variational inference problems, 2026. URL https://arxiv.org/abs/2606.25492.

Blaise Pascal and Pierre de Fermat. Correspondence on the problem of points and games of chance. Letters; edited English resource hosted by the University of York, 1654. URL https://www.york.ac.uk/depts/maths/histstat/pascal.pdf.

Roger D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011. doi: 10.1126/science.1213847.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154, 2022. doi: 10.1109/TSP.2022.3153135. URL https://doi.org/10.1109/TSP.2022.3153135.

Nicolas P. Rougier, Michael Droettboom, and Philip E. Bourne. Ten simple rules for better figures. *PLOS Computational Biology*, 10(9):e1003833, 2014. doi: 10.1371/journal.pcbi.1003833. URL https://doi.org/10.1371/journal.pcbi.1003833.

Ryan Smith, Philipp Schwartenbeck, Thomas Parr, and Karl J. Friston. Active inference, bayesian optimal design, and expected utility. In *The Drive for Knowledge: The Science of Human Information Seeking*. Cambridge University Press, 2022. doi: 10.1017/9781009026949.007.

Volker Tresp. A bayesian committee machine. *Neural Computation*, 12(11):2719–2741, 2000. doi: 10.1162/089976600300014908. URL https://www.dbs.ifi.lmu.de/~tresp/papers/bcm6.pdf.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2): 129–133, 2016. doi: 10.1080/00031305.2016.1154108. URL https://doi.org/10.1080/00031305.2016.1154108.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://arxiv.org/abs/1805.07836.