

Active Inference as a Meta-Pragmatic Framework

Executable Models, Evidence, and the Discipline of Revision

Daniel Ari Friedman

Active Inference Institute, FractiAI

ORCID: [0000-0001-6232-9096](https://orcid.org/0000-0001-6232-9096)

DOI: [10.5281/zenodo.23397453](https://doi.org/10.5281/zenodo.23397453)

October 15, 2026

FOUR ROLES IN META-PRAGMATIC INQUIRY

Q = quadrant; Q1-Q4 name overlapping inquiry roles, not stages.

Q1 Infer and act

Update beliefs from observations;
evaluate actions under a model.

Q2 Contextualize evidence

Condition on source reliability;
examine the context of evidence.

Q3 Monitor inquiry

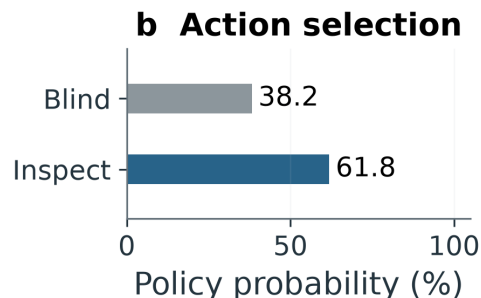
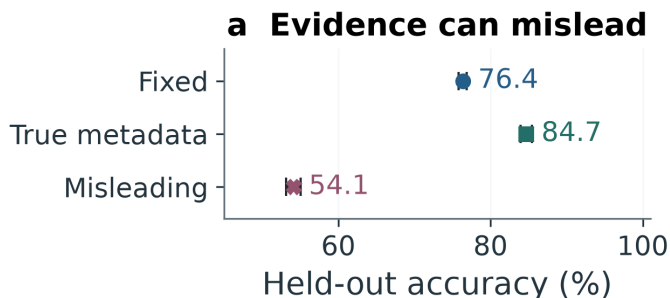
Assess uncertainty and cost;
regulate further observation.

Q4 Evaluate and revise

Challenge model assumptions;
justify revisions before adoption.

GNN = Generalized Notation Notation: executable model specifications.

Model execution → evaluated consequences → researcher-led revision



Sensor points: seed means and marginal 95% intervals.

Action bars: exact policy from expected free energy. Both tasks are synthetic.

Contents

1	Abstract	3
2	The Representational Commitments That Make Inference Possible	4
2.1	Computational, methodological, and public responsibility	5
3	American Pragmatism and the Public Criticism of Models	5
3.1	Peirce: doubt, habit, and public correction	5
3.2	James: verification, consequences, and the correction of truth claims	5
3.3	Dewey: reflective inquiry and the reconstruction of ends	6
3.4	Addams: whose experience can correct the evaluator?	6
3.5	Mead: significant symbols and meaning through shared conduct	6
3.6	What the pragmatist inheritance asks of an executable model	6
3.7	Active inference, generative models, and explicit specification	6
3.8	OODA, active inference, and criticism across the action loop	6
3.9	Metacognitive monitoring, control, and forecast calibration	7
3.10	Outcome preferences, reward representations, and action selection	7
3.11	AI safety, information integrity, and evidential vulnerability	7
3.12	Bayesian model criticism and the resource limits of inquiry	7
3.13	Out-of-sample predictive assessment and the meanings of calibration	7
3.14	Resources and the scope of sensitivity analysis	8
3.15	Situated inquiry and the origins of commitments	8
3.16	Research economy and public representations	8
3.17	Executable specification as an inspectable scientific artifact	9
4	Executable Generative Models of Evidence, Decisions, and Action	9
4.1	Two experimental models and their distinct explanatory questions	9
4.2	Executable declarations make modeling commitments inspectable	10
5	Testing Model Commitments Through Controlled Synthetic Studies	11
5.1	Source reliability, evidence acquisition, and supervised learning	11
5.2	Action selection under explicit preferences and model assumptions	12
5.3	Experimental units, uncertainty, and the scope of the evidence	13
6	Four Complementary Roles in Meta-Pragmatic Inquiry	13
6.1	The two conceptual axes and four operational responsibilities	13
6.2	Q1: infer hidden states and evaluate actions within a model	13
6.3	Q2: contextualize evidence through source reliability	13
6.4	Q3: monitor uncertainty and regulate evidence acquisition	13
6.5	Q4: evaluate model assumptions and justify their revision	14
6.6	Overlapping roles, comparative predictions, and falsifiability	14
7	Evidence Integrity, Source Trust, and Model Vulnerability	14
7.1	A scoped threat model for evidence and model commitments	14
7.2	Misleading metadata as a measured synthetic failure case	14
7.3	Trust, calibration, and the limits of diagnosis	14
7.4	Bounded integrity safeguards and unresolved assurance questions	14
7.5	The separate claims an assurance argument must establish	15
8	Interpreting Experimental Evidence and Justifying Model Revision	15
8.1	What the controlled experiments establish about these mechanisms	15
8.2	What the four-role taxonomy contributes to model criticism	15
8.3	Meta-epistemic practice as fallible inquiry	15
8.4	Meta-pragmatic practice and the authority to revise	15
8.5	Comparing prediction quality, sensing resources, and calibration	16
8.6	Stopping inquiry requires an explicit resource-sensitive judgment	17
8.7	Action selection reveals the preferences and assumptions of a model	17
8.8	Exact inference and confident error in a misspecified world	18
8.9	Limitations of the synthetic design and philosophical interpretation	19
8.10	From controlled synthetic examples to independent empirical tests	20

9	Toward Accountable Active Inference Through Public Criticism	20
10	Acknowledgments, Author Identity, and Research Integrity	21
10.1	Author identity and the versioned preprint archive	21
10.2	Synthetic intelligence assistance and human accountability	21
10.3	Synthetic data, retained artifacts, and research declarations	21
11	Technical Supplement: Formal Models and Statistical Protocols	21
11.1	Imported GNN tensors behind the experimental model diagrams	22
11.2	Categorical probability contracts and exact Bayesian inference	23
11.3	Complete synthetic experimental design and statistical protocol	24
11.4	Paired contrasts, stopping rules, and policy surfaces	27
11.5	Model revision mechanics and the limits of integrity diagnostics	31
11.6	Derivations, retained results, and reproduction	32
12	Supplemental Notation for Models, Objectives, and Statistics	33
12.1	States, observations, and executable model components	34
12.2	Information measures, evaluative objectives, and policy selection	35
12.3	Statistical estimands, uncertainty, and computational notation	36
12.4	Conceptual terms and epistemic boundaries	36
13	References	36

1 Abstract

Active inference makes beliefs and policies conditional on a generative model, leaving open how its categories, evidence and purposes become warranted. We develop active inference as a meta-pragmatic framework for making these conditions of inquiry themselves available to inquiry. Four overlapping roles organize the proposal: inferring states and evaluating actions, contextualizing evidence, monitoring uncertainty and resources, and evaluating or revising model commitments. These roles distinguish responsibilities within inquiry rather than successive stages or separate cognitive faculties. In dialogue with American Pragmatism, situated cognition and Bayesian model criticism, the framework treats knowledge as answerable to consequences while keeping predictive success, explanatory understanding and the justification of ends distinct. Mathematical coherence cannot establish the adequacy of a representation, and optimization under declared preferences cannot justify those preferences. Public criticism consequently concerns both the grounds of a decision and whose experience can challenge the purposes it serves.

The proposal is made concrete through executable Generalized Notation Notation (GNN) specifications imported with the up-stream reference parser and evaluated by an exact categorical runtime. Complementary synthetic studies examine source reliability, misleading metadata, confidence-triggered evidence acquisition, likelihoods fitted on separate labeled training data, and action selection by expected free energy. Paired comparisons across independent simulation replicates retain predictive scores, observation costs and uncertainty. Reliable metadata and further sensing improve classification in the specified sensor task, while misleading metadata and a separately specified world reveal how exact inference can remain confidently mistaken. The contribution is an inspectable discipline of specification, evaluation and accountable revision, supported by model diagrams, retained evidence and technical supplements. The demonstrations establish conditional computational consequences; they neither automate the justification of ends nor validate the framework’s broader biological or social interpretation. Sources, model specifications and retained evidence accompany the manuscript at https://github.com/docxology/active_inference_meta_pragmatic.

Keywords: active inference; meta-pragmatism; inquiry; model criticism; evidence; generative models

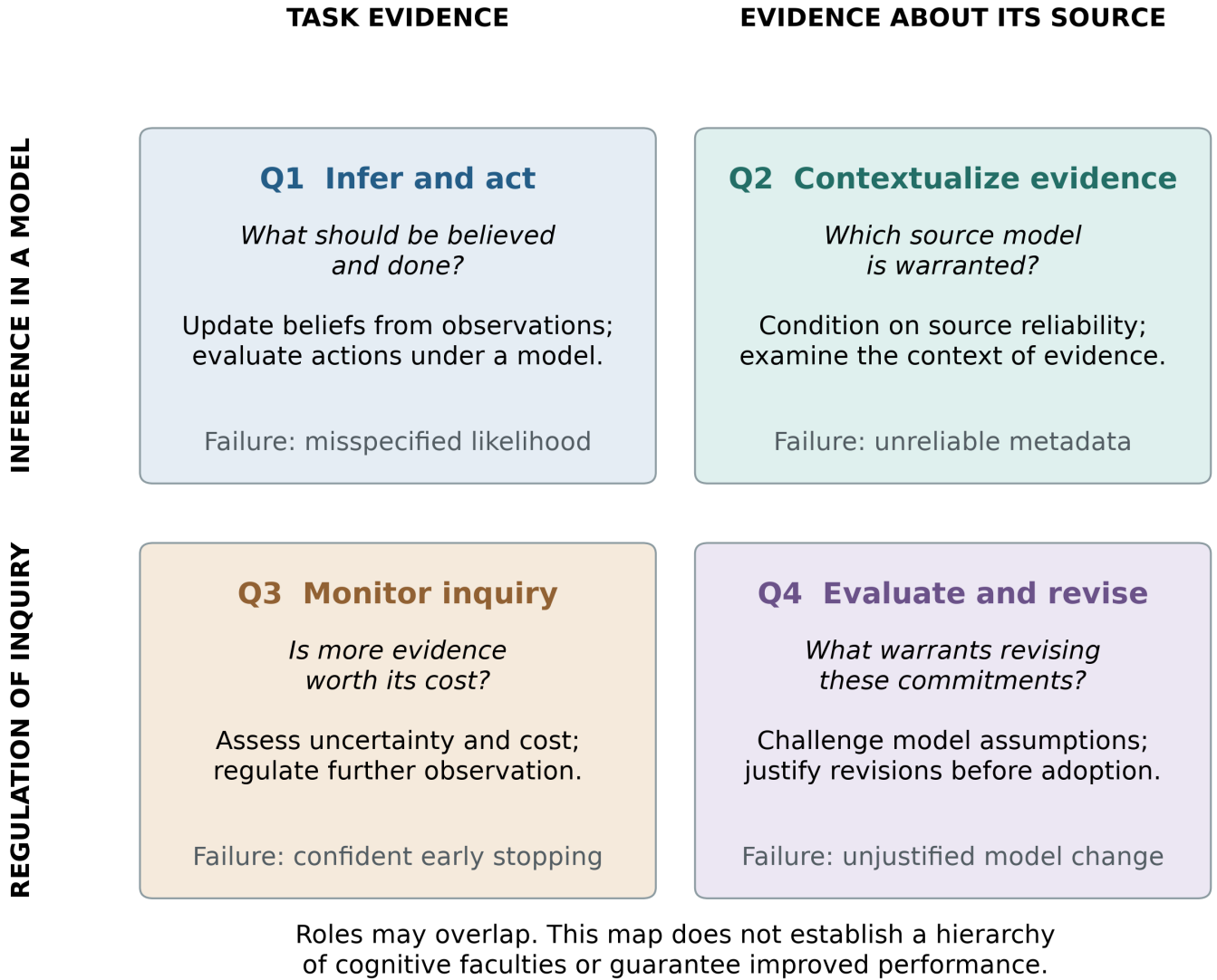


Figure 1: Four complementary roles in meta-pragmatic inquiry. Q denotes quadrant: Q1 infers states and evaluates actions; Q2 contextualizes evidence; Q3 monitors uncertainty and regulates acquisition; Q4 evaluates assumptions and justifies revision. Columns distinguish task evidence from information about its source, and rows distinguish inference within a model from regulation of inquiry. Each cell states a guiding question, operation and failure case. These overlapping responsibilities are neither sequential stages nor measured cognitive faculties; no historical pragmatist is assigned to a single quadrant.

2 The Representational Commitments That Make Inference Possible

An agent can infer correctly under an inappropriate observation model or select the minimum-cost policy under inappropriate preferences. In either case, improving the inference algorithm alone may leave the error intact. This distinction motivates examining how a model is specified, evaluated, and revised.

We use *meta-epistemic* for choices about the representation and acquisition of evidence, and *meta-pragmatic* for choices about preferences and the evaluation of action. These are interpretive terms for modeling activities. They do not imply that ordinary parameter estimation is impossible or that an agent’s initial ontology is irrevocable. Observation and transition models can be learned; the selection of representations, update rules, and constraints remains a modeling decision.

Figure 1 gives the four roles a shared visual reference before their formal development. Our contribution is a taxonomy connecting these activities to inspectable computational mechanisms. We distinguish data from information about its source or reliability, and direct inference from evaluation of inference. The resulting quadrants organize examples and failure cases. Their usefulness depends on operational definitions and comparative tests; the taxonomy itself does not guarantee improved performance. A model is therefore both a device for inference and an object of criticism: its answers can be internally consistent while its assumptions remain inadequate.

The paper contributes (1) explicit definitions of the four quadrants, (2) executable Generalized Notation Notation (GNN) specifications imported through the upstream reference parser, (3) an exact categorical runtime with independent mathematical checks, (4) paired synthetic comparisons that retain evidence and account for sensing resources, and (5) a philosophical account of model criticism and responsible revision. Active inference is not the only framework in which these questions arise. Reinforcement learning, Bayesian decision theory, active learning, and meta-learning also permit learned models, structured objectives, and information-seeking behavior.

2.1 Computational, methodological, and public responsibility

We call active inference a *meta-pragmatic framework* when its explicit models support inquiry into the consequences of modeling and acting: which distinctions are represented, which evidence is acquired, which outcomes are preferred, and what would justify revising those commitments. The framework consists of a declared model, an executable decision procedure, an external evaluation reference, and an accountable revision process. Only the first two are supplied by an inference calculation. Meta-pragmatic practice connects all four; it is not an additional hidden optimization objective that automatically establishes the authority of its own preferences.

This interpretation follows a practical-consequence tradition while preserving differences between philosophical explanation and computational evidence. Peirce connects inquiry to unsettled belief and conceptual meaning to conceivable practical effects Peirce [1877, 1878]. We operationalize that orientation through interventions on model components and their evaluated consequences. Our synthetic experiments test those interventions in declared worlds; they do not demonstrate that the philosophical interpretation is uniquely correct.

At the **computational level**, a declared model determines what counts as evidence and which policies receive favorable scores. At the **methodological level**, a researcher chooses comparisons, scoring rules, budgets, and circumstances under which a model will be revised. At the **normative level**, someone must justify which outcomes matter and whose interests constrain revision. A successful calculation at one level does not discharge obligations at the others. Calling a process meta-epistemic or meta-pragmatic is useful only when its referent and evaluation criterion are stated.

The executable specification is the bridge between the first two levels. GNN exposes likelihoods, transitions, preference logits, priors, dimensions, and dependency declarations before execution. It makes a model inspectable and portable within a supported profile; it does not establish that the model is true or its preferences legitimate. The present implementation imports real upstream code rather than treating a diagram or a notation label as an executed model.

Section 3 places the interpretation in its research context. Section 4 makes the executed models visible. Sections 5 and 6 give the experimental design and operational taxonomy. Sections 7, 8, and 9 delimit the implications. The technical supplement retains derivations, full protocols, sensitivity analyses and reproducibility details; Supplement 12 gathers the notation. Retained artifacts provide the audit trail.

3 American Pragmatism and the Public Criticism of Models

American Pragmatism asks how inquiry earns and revises its commitments through consequences, criticism and shared experience. Its plural traditions supply distinct questions for this framework, rather than a historical derivation of the four roles or an anticipation of categorical active inference.

3.1 Peirce: doubt, habit, and public correction

Peirce distinguishes scientific inquiry, answerable to an independent reality, from tenacity, authority and the a priori method Peirce [1877]. These four methods concern the settlement of belief; they do not correspond to our four intersecting roles. His pragmatic maxim clarifies a conception through conceivable practical consequences Peirce [1878]; his later account connects belief, habit, surprise and self-control Peirce [1905]. The modeling question is what a commitment leads us to expect and what could unsettle it. Posterior concentration alone supplies neither genuine doubt nor public correction: our misleading-metadata control shows internally coherent inference under inadequate premises.

3.2 James: verification, consequences, and the correction of truth claims

James asks what difference competing ideas make in experience, while making verification, consistency and subsequent correction central to truth claims James [1907], Lectures II and VI. This directs attention to predicted observations, changed actions, incurred costs and possible counterevidence. It does not identify truth with convenience or one performance score. Accuracy and proper losses reveal different consequences; their selection and evaluation horizon remain contestable commitments.

3.3 Dewey: reflective inquiry and the reconstruction of ends

For Dewey, reflection examines the grounds and consequences of belief and sustains inquiry through a problematic situation Dewey [1910], Chapter I. A confidence-triggered observation rule captures only a narrow decision about continuing investigation. His treatment of ends through needs, opportunities and consequences also makes purposes available for reconstruction Dewey [1920], Chapters VI–VII. Computing an action under C therefore differs from asking whether C expresses an adequate end. Our fixed evaluation criteria expose this distinction without implementing autonomous reconstruction of purposes.

3.4 Addams: whose experience can correct the evaluator?

Addams challenges the sufficiency of private virtue for social ethics and brings diverse experience into the correction of moral judgment Addams [1902], Introduction and Charitable Effort. Her question for evaluation is whose experience can contest the declared account of benefit and harm. A held-out test can assess a chosen outcome while excluding affected perspectives from its definition. Our synthetic contexts contain no social participants; participatory criticism is a normative requirement proposed for social applications, not a demonstrated software capability.

3.5 Mead: significant symbols and meaning through shared conduct

Mead situates significant symbols in coordinated social conduct and the indication of anticipated consequences Mead [1934], Sections 16 and 24 of the posthumous volume edited by Charles W. Morris. A GNN label declares an intended meaning without establishing shared interpretation. Executable specifications can make disagreement about categories precise, but our single-agent tasks do not implement symbolic interaction or the social development of a self.

3.6 What the pragmatist inheritance asks of an executable model

Together these readings ask us to articulate consequences, expose reasons, and permit criticism of representations and ends. The four roles cut across those responsibilities; no author belongs to one quadrant. In the demonstrations, the researcher specifies the world, model and evaluation. Public criticism and the social formation of ends remain outside the mechanism. A meta-pragmatic framework makes these choices inspectable without automating their justification or substituting philosophical authority for held-out evidence.

3.7 Active inference, generative models, and explicit specification

The free energy principle and discrete active inference supply the background for the distinction between state inference and model specification Friston [2010], Da Costa et al. [2020]. Expected free energy balances information gain and outcome preferences under a predictive model Friston et al. [2015]. The variational assumptions behind planning require care: expected free energy is not obtained by simply substituting future observations into the usual variational objective Millidge et al. [2021]. Our implementation is deliberately limited to categorical models with known likelihood and transition matrices.

Methods and software already address learning, hierarchical models, and practical computation Friston et al. [2016], Smith et al. [2022], Heins et al. [2022]. Thus, our meta-level interpretation does not introduce model learning to active inference. Its contribution is a classification of specification and monitoring choices, coupled to explicit examples that can be checked computationally.

Recent work studies scale-free generative models, variational planning, and interpretable graph-based agents Friston et al. [2025], de Vries et al. [2025], Pazem et al. [2025]. These are distinct technical architectures, not implementations or validation of the quadrants proposed here. We include them to clarify the surrounding research program, without importing their performance claims into our results.

3.8 OODA, active inference, and criticism across the action loop

Boyd’s observe–orient–decide–act (OODA) framework provides a useful comparison for inquiry organized around action. David, Cordes and Friedman explicitly connect OODA with active inference in *Active Inference in Modeling Conflict* David et al. [2021], especially pp. 17–18 and 37–38, Figure 9. Their alignment relates observation to incoming signals, orientation to inference about hidden causes, decision to preference-sensitive policy evaluation, and action to an enacted policy. This is a conceptual antecedent, not empirical validation of our taxonomy or experiments.

The following crosswalk is our interpretation of the present executable tasks. It distinguishes a recurring action loop from responsibilities that can operate throughout it.

OODA function	Connection to the present framework
Observe: receive evidence	Q1 updates from observations; Q3 controls whether further evidence is acquired.
Orient: interpret the situation	Q2 contextualizes evidence through source reliability; Q4 evaluates the assumed reliability model.
Decide: evaluate possible conduct	Q1 evaluates policies under declared preferences; Q3 exposes sensing costs and stopping criteria.
Act: intervene and encounter consequences	The action model’s transition changes observation mode; external evaluation can then inform Q4 criticism.

The shared count of four implies no one-to-one mapping. Repeatedly executing OODA under false source assumptions can preserve a bad orientation; the misleading-metadata and world-agent mismatch controls make narrower versions of this vulnerability inspectable. Our one-step action task executes a bounded perception-action coupling, not a strategic conflict simulation. The meta-pragmatic question cuts across the loop: what evidence, representation, purpose or resource judgment warrants revision, and who may authorize it?

3.9 Metacognitive monitoring, control, and forecast calibration

The distinction between monitoring and control predates this taxonomy [Flavell \[1979\]](#), [Nelson and Narens \[1990\]](#). Confidence, calibration, discrimination, and metacognitive efficiency are different quantities [Fleming and Lau \[2014\]](#). A concentrated posterior can be confidently wrong under model misspecification. Accordingly, the experiment reports Brier loss and log loss rather than treating posterior concentration as evidence of correctness.

MetaMind and SOFAI provide contemporary examples of metacognitive organization in artificial systems [Zhang et al. \[2025\]](#), [Bergamaschi Ganapini et al. \[2025\]](#). Their tasks and mechanisms differ from our binary sensor setting. An architectural analogy does not establish shared statistical behavior or biological equivalence.

3.10 Outcome preferences, reward representations, and action selection

The difference between specifying outcome preferences and specifying rewards is useful, but it is not an absolute division between inspectable and opaque objectives. If preferences are encoded as $p_C(o) \propto \exp(\beta r(o))$, the expected log-preference term corresponds to expected reward up to a scale and normalization constant. Epistemic value adds a particular information-seeking objective. Active inference and model-based reinforcement learning have operational connections [Tschantz et al. \[2019\]](#), [Sajid et al. \[2021\]](#); both require assumptions about observations, actions, and values.

Our question is therefore comparative: which explicit reliability or monitoring mechanism changes behavior under a specified environment and resource budget? The four labels do not determine the answer in advance.

3.11 AI safety, information integrity, and evidential vulnerability

Objective specification, distribution shift, and oversight are established concerns in AI safety [Amodei et al. \[2016\]](#), [Gabriel \[2020\]](#). We use the quadrants to organize where an intervention acts: on observations, metadata, monitoring rules, or model revision. We do not infer human susceptibility, attack success probabilities, or alignment guarantees from the taxonomy. The misleading-metadata control is a controlled synthetic failure case, not a deployment security evaluation.

3.12 Bayesian model criticism and the resource limits of inquiry

Gelman and Shalizi distinguish inference within a model from criticism and revision of the model itself [Gelman and Shalizi \[2013\]](#). That distinction motivates our separation of posterior updating from the evaluation of its assumptions. Simon’s bounded account of rational choice makes resource limitations relevant to a decision procedure [Simon \[1955\]](#); accordingly, an accuracy improvement obtained with extra observations is evaluated alongside its cost.

Proper scoring rules provide a principled basis for assessing probabilistic forecasts [Gneiting and Raftery \[2007\]](#). We use Brier and logarithmic losses in addition to classification accuracy. These losses assess forecast quality; neither is a pure calibration statistic. Descriptive calibration diagrams separately compare forecast probabilities with held-out frequencies, following the distinction between coherence and calibration [Dawid \[1982\]](#).

3.13 Out-of-sample predictive assessment and the meanings of calibration

Bayesian workflow treats model construction, checking, computation, and comparison as linked activities rather than reducing scientific practice to a posterior calculation [Gelman et al. \[2020\]](#). This supplies a methodological neighbor for the four roles: each

role identifies a place where inquiry can change course, while workflow specifies how a sequence of such changes is evaluated. Our synthetic study implements one held-out evaluation cycle; it does not establish convergence of an open-ended revision process.

Predictive evaluation must follow the sampling structure. Demšar’s comparisons concern algorithms across multiple datasets [Demšar \[2006\]](#); our seed replicates are generated datasets within one environment, not independently chosen scientific domains. Thus, seed uncertainty cannot stand in for cross-domain generalization. Vehtari and colleagues analyze out-of-sample predictive evaluation with LOO and WAIC [Vehtari et al. \[2017\]](#). We use a separately generated test set and frozen fitted reliabilities; we do not claim to implement those estimators or their diagnostics.

The word calibration also needs an explicit object. Guo and colleagues examine confidence calibration in neural classifiers [Guo et al. \[2017\]](#), whereas our curves assess probabilities of state 1 in a binary generative task. Vaicenavicius and colleagues give a broader framework for evaluating probabilistic calibration [Vaicenavicius et al. \[2019\]](#). Their distinctions motivate treating a binned reliability curve as a diagnostic with finite-sample and binning limitations. Simulation-based calibration instead checks an inference algorithm across generative draws [Talts et al. \[2018\]](#). Our algebraic VFE/EFE identities and held-out forecast curves are different checks and should not be relabeled as simulation-based calibration.

3.14 Resources and the scope of sensitivity analysis

Gigerenzer and Gaissmaier’s ecological approach to heuristics makes strategy quality conditional on the structure of the decision environment [Gigerenzer and Gaissmaier \[2011\]](#). This sharpens the interpretation of our stopping rule: a rule can economize on observations in this task without becoming a generally rational procedure. Our comparison does not establish which heuristics people use, nor does it estimate an ecological optimum.

Resource-rational analysis asks how computational constraints enter an account of cognitive performance [Lieder and Griffiths \[2020\]](#). Our observation counts operationalize one resource; they do not measure cognitive effort, time, or energy. The comparison therefore supports a resource accounting argument while leaving the mapping to biological mechanisms unresolved.

Steegeen and colleagues make analytical alternatives visible through multiverse analysis [Steegeen et al. \[2016\]](#). Our threshold sweep follows that transparency motivation in a narrower setting: eleven explicit acquisition thresholds reuse the same test trials. It is a sensitivity analysis, not an exhaustive multiverse or an independent validation of a selected policy. Any threshold chosen from this sweep would need a new evaluation set.

3.15 Situated inquiry and the origins of commitments

Dewey’s treatment of the reflex arc emphasizes coordinated activity rather than an isolated stimulus followed by an isolated response [Dewey \[1896\]](#). This motivates asking how observation and action jointly define a task. Our action model implements one small, explicit coupling: an intervention changes the observation mode while preserving the hidden context. It does not reproduce Dewey’s account of organismic coordination.

Friedman’s *Before Pragmatism Had a Name* develops a reading of Blake alongside pragmatist inquiry and active inference [Friedman \[2026a\]](#). We draw a bounded methodological lesson from this conceptual juxtaposition: an inherited representational habit can itself become an object of inquiry. The historical and literary synthesis is an attributed interpretation, not empirical support for the present taxonomy or a mathematical equivalence between Blake’s figures and generative-model variables.

Synthesis of Agent and Niche stages a speculative encounter between Blake and ecological accounts of cognition [Friedman \[2025\]](#). It directs attention toward the setting that makes an observation meaningful. Here that attention becomes a testable modeling distinction between a declared agent and a separately specified world. The literary work supplies conceptual context, whereas the retained executable specifications and scores supply evidence for the numerical comparison.

Prior Cognitive Art asks which explanatory question is being answered when the origin or persistence of a prior is invoked [Friedman \[2026b\]](#). Its conceptual organization draws on Tinbergen’s questions about causation, development, survival value, and evolution [Tinbergen \[1963\]](#). A fitted probability, a developmental account, and an evolutionary explanation are therefore different explanatory objects. Our fixed categorical prior is an input to a calculation. Neither a successful posterior update nor the addition of a hyperprior explains its biological origin. The cited working paper presents conceptual and deterministic illustrations, not stochastic performance evidence for our study.

3.16 Research economy and public representations

Beni and Pietarinen connect active inference with Peirce’s abduction, induction and economy of research [Beni and Pietarinen \[2021\]](#). Their philosophical alignment is a close antecedent to the present interpretation. Our narrower contribution is to expose concrete commitments and their evaluation through retained executable models. Selecting an informative action within a fixed representation is not equivalent to inventing a representation; reducing prediction error alone does not authorize an unrestricted change of ends.

Diagrams also participate in scientific accountability. A visualization can make assumptions, units and comparisons inspectable, or conceal them through scale, typography and omission. Rougier and colleagues treat figure design as communication between

data and readers Rougier et al. [2014]. Here the conceptual quadrant map, exact model surfaces and seed-based estimates have different captions and visual roles. Readability does not turn a conceptual diagram into evidence, and numerical precision does not make a model adequate. A public representation should expose the distinction.

3.17 Executable specification as an inspectable scientific artifact

The GNN project defines a structured notation and parsing pipeline for active-inference generative models Active Inference Institute [2026]. We import its reference Markdown parser at a recorded revision and consume the resulting parameter values. Our adapter admits a bounded, single-factor categorical profile. GNN is the specification dependency; the local exact categorical equations are the execution engine. No claim is made that every upstream backend, model family, or parser acceptance condition is supported by this adapter.

4 Executable Generative Models of Evidence, Decisions, and Action

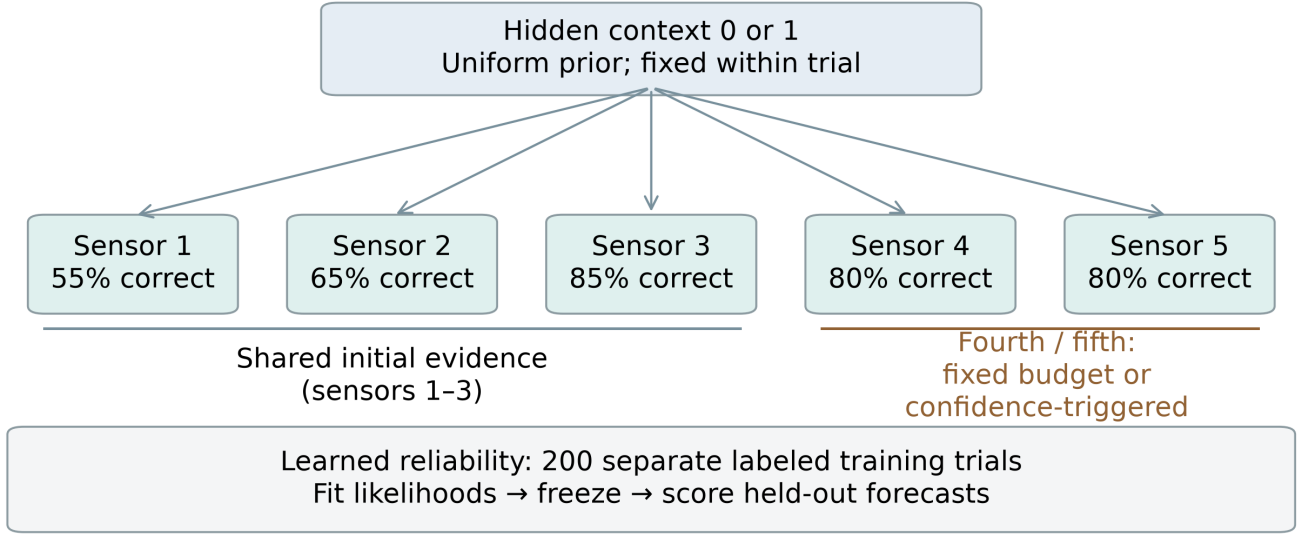
A generative model expresses how hidden conditions could produce observations and how actions could change what happens next. Active inference uses that model both to interpret evidence and to evaluate possible action. The crucial distinction is between a world that generates an observation and an agent’s assumptions about that world. An agent can execute its assumptions precisely while misunderstanding the reliability of what it observes.

4.1 Two experimental models and their distinct explanatory questions

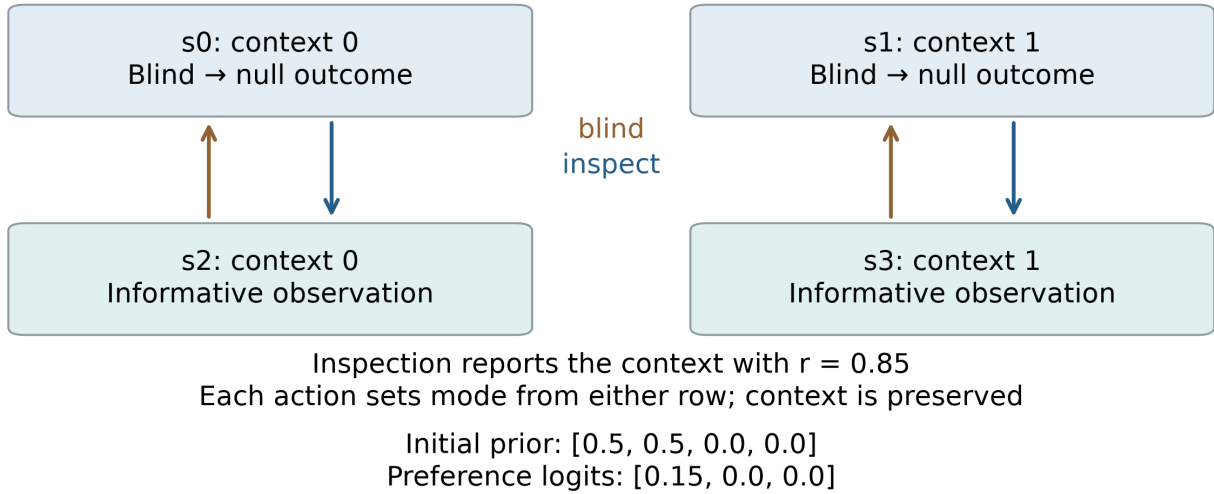
Figure 2 separates the two experimental settings. The sensor study asks how beliefs change when the reliability of an observation is ignored, correctly supplied, learned or falsely reported. Its hidden context remains fixed during a trial. Several sensors offer conditionally independent views of that context; acquiring another view consumes an explicit resource. A confidence-triggered acquisition rule is a monitoring mechanism, not an EFE policy.

The action study asks a different question: when does an agent choose an informative observation? Its hidden state combines context with observation mode. Actions preserve context while choosing between a blind mode, which returns a null observation, and an informative mode, which provides a noisy indication of context. Expected free energy evaluates the consequences of these actions under declared preferences as well as their information value. Neither the preference specification nor a successful calculation establishes which ends should be adopted.

a Sensor study: reliability and evidence acquisition



b Action study: interventions change observation mode



A: observation likelihoods · B: action-dependent state transitions
C: preference logits · D: initial beliefs · External labels score forecasts.

Figure 2: Two executable models and the evidence each can supply. (a) A shared binary context generates five potential sensor reports; percentages give each sensor’s probability of reporting the true context. All strategies begin with sensors 1–3. Additional observations follow a fixed budget or a confidence rule; fitted reliability uses separate labeled training. (b) The imported GNN action model preserves context while choosing blind or informative observation mode. Arrows show nonself transitions, with self transitions implicit. Likelihood A, transitions B, preference logits C and initial beliefs D define the agent; external context labels assess forecasts rather than supplying the agent’s confidence.

4.2 Executable declarations make modeling commitments inspectable

The models declare observation likelihoods, action-dependent transitions, preferences and initial beliefs in real GNN specifications. The upstream reference parser supplies the tensors consumed by the local categorical runtime. This connects a readable declaration to the calculation actually performed; it does not certify the declaration’s truth. The supplementary tensor atlas, Figure 7, also shows the packaged stay/switch control model used for numerical checks. It is distinct from the two retained experimental tasks.

Bayesian updating revises beliefs conditional on a model. Expected free energy evaluates prospective information and preferred outcomes within that model [Da Costa et al. \[2020\]](#), [Friston et al. \[2015\]](#). Criticism asks whether the model deserves continued use [Gelman and Shalizi \[2013\]](#). Keeping these activities distinct prevents an exact posterior from becoming a certificate of adequate representation. The equations, probability conventions, numerical contracts and implementation boundaries are collected in Supplement [11.2](#); the complete notation guide is in Supplement [12](#).

5 Testing Model Commitments Through Controlled Synthetic Studies

The framework becomes scientifically useful when its roles identify interventions with observable consequences. The studies therefore compare concrete mechanisms rather than treating the quadrant names as algorithms. Their design is synthetic, fixed in code and not preregistered. No participant data or deployment outcomes are analyzed. Complete task parameters, randomization, statistical procedures and retained-file definitions are in Supplement [11.3](#).

5.1 Source reliability, evidence acquisition, and supervised learning

The sensor study compares a fixed reliability assumption with source-specific metadata, confidence-triggered acquisition and reliability learned from separate labeled training. Every strategy shares the same held-out contexts and potential sensor observations within each seed. A misleading-metadata control tests whether contextual information can make inference worse. Fixed sensing budgets distinguish the effects of additional evidence from those of a stopping rule.

Figure [3](#) shows why a single ranking would obscure the argument. Reliable metadata improves a misspecified baseline in this environment; further evidence can improve predictions while consuming more observations. Fitting reliability is an externally supervised intervention and carries a training cost. Falsely reported metadata can undermine a mathematically valid update. These are consequences of explicit model choices, not evidence that higher-order processing is intrinsically superior.

Synthetic binary sensor task

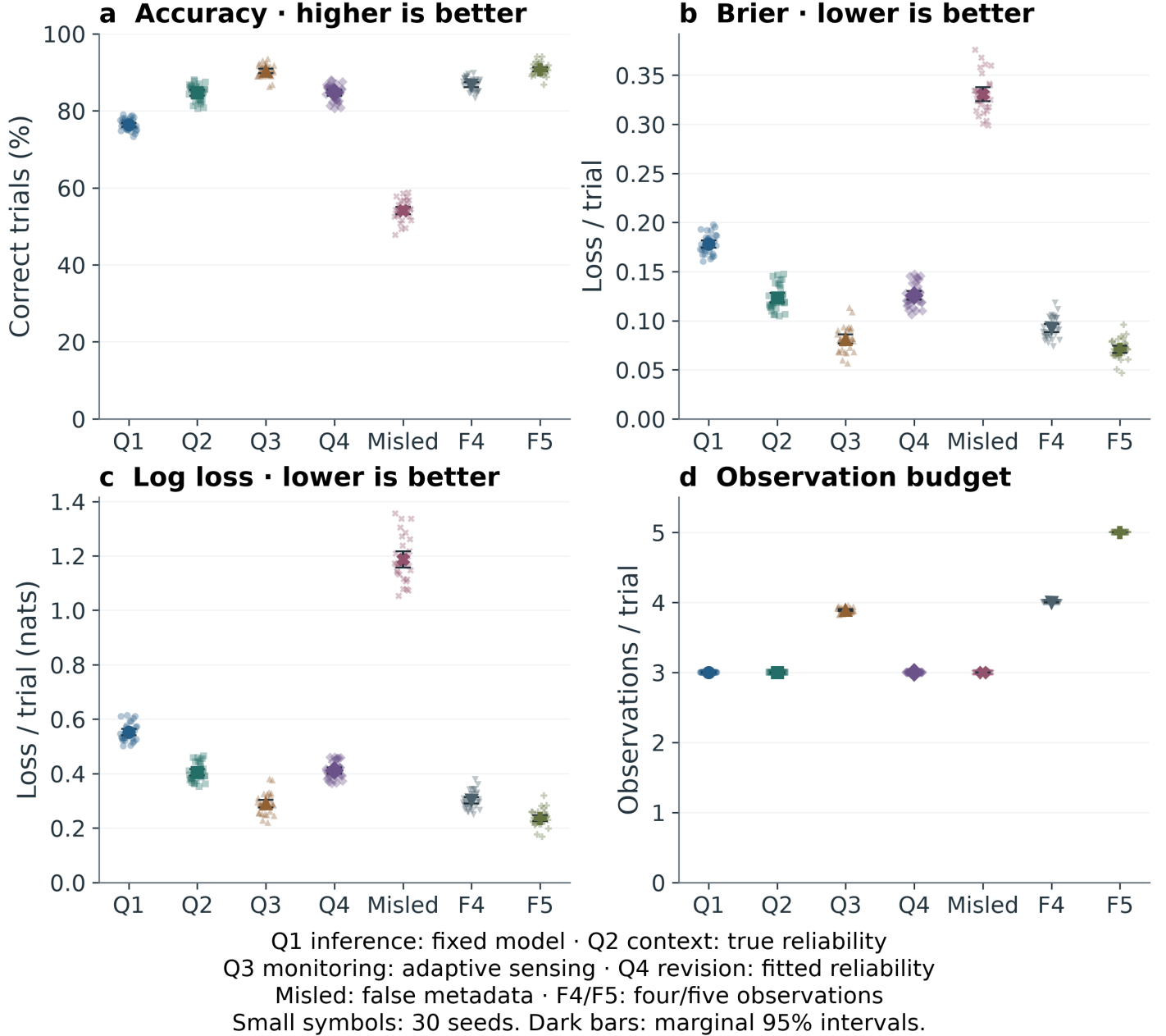


Figure 3: Predictive performance and observation use under seven explicit sensor strategies. Small symbols show 30 independent seed means; large symbols and capped dark bars show means and marginal 95 percent seed-bootstrap intervals. (a) Higher accuracy is better; (b,c) lower Brier and logarithmic loss is better; (d) acquired observations expose resource differences. Q1–Q4 refer to the implemented fixed-model, true-metadata, adaptive-sensing and fitted-model conditions, respectively; they do not test the full philosophical roles. Misled uses false metadata; F4/F5 use four/five reports. Counts exclude separate training. Paired contrasts, rather than overlap of these marginal intervals, address strategy differences.

5.2 Action selection under explicit preferences and model assumptions

The separate action task samples the policy probabilities actually returned by EFE evaluation. Always-inspect and always-blind baselines share the generated contexts and potential informative outcomes. Preferences can make an uninformative action attractive; a policy can therefore sacrifice classification accuracy without violating its declared objective. The mechanism is inspected as an example of how information and value enter action, rather than as a universal prescription for sensing.

An exact comparison then fixes the agent’s assumptions while varying the world’s actual reliability. Each side has a separately imported GNN specification. Forecasts are evaluated using world probabilities, so a confident but mistaken likelihood becomes visible without attributing Monte Carlo noise to the discrepancy. This separation is essential to model criticism: the model’s own expectations cannot be its only tribunal.

5.3 Experimental units, uncertainty, and the scope of the evidence

Independent generated seeds are the unit of uncertainty. Paired seed estimates, proper scoring losses, observation counts and prespecified adverse controls provide complementary views of each mechanism. Training data remain separate from held-out evaluation. Bootstrap intervals are marginal; the primary sign-flip tests retain their stated exchangeability assumption and three-comparison Holm family [Efron and Tibshirani \[1993\]](#), [Holm \[1979\]](#). Exploratory sensitivity analyses introduce no new confirmatory rejection claims.

The exact reliability grid is finite enumeration, not a collection of empirical replications. Its curves receive neither sampling intervals nor p-values. Probabilistic scores assess forecast quality, while calibration examines agreement between forecasts and outcomes [Gneiting and Raftery \[2007\]](#), [Dawid \[1982\]](#). Computational checks, synthetic behavior and philosophical interpretation support different conclusions; none substitutes for the others.

6 Four Complementary Roles in Meta-Pragmatic Inquiry

6.1 The two conceptual axes and four operational responsibilities

The horizontal axis distinguishes observations from information about their production, reliability, or provenance. The vertical axis distinguishes performing inference from monitoring, evaluating, or revising inference. A datum can be metadata relative to one task and ordinary data relative to another; classification must therefore specify the object-level task and monitored process.

	Data	Metadata
Object-level cognition	Q1: infer states and evaluate actions	Q2: condition inference on source reliability or context
Meta-cognitive regulation	Q3: monitor uncertainty and alter processing or acquisition	Q4: evaluate or revise the reliability model, strategy, or representation

The axes are analytically distinguishable, but we do not assert statistical independence or universal neural separability. One implementation can contain operations assigned to several quadrants. Neither a timestamp nor a confidence value alone makes a system meta-cognitive; there must be a specified monitoring/control relation. The OODA comparison in [Section 3.8](#) makes the same distinction: these responsibilities cut across a perception–action loop rather than naming four successive stages.

The early map in [Figure 1](#) provides the shared orientation for these roles. A complementary revision-cycle diagram appears in [Supplement 11.5](#), [Figure 12](#).

6.2 Q1: infer hidden states and evaluate actions within a model

Q1 includes [Equation \(1\)](#) and policy evaluation in [Equation \(3\)](#) under a fixed model. The experimental Q1 baseline uses the same inference rule as the other conditions, with a homogeneous reliability assumption. Its weakness is model misspecification, rather than the absence of any Bayesian operation. Correctly specified Q1 inference can match a metadata-aware strategy when metadata carries no additional information.

6.3 Q2: contextualize evidence through source reliability

Q2 uses metadata m by conditioning the likelihood, $p(o | s, m)$. In the sensor example, m identifies sensor-specific reliability. This changes the evidence supplied to the same Bayesian update. The gain is conditional on the metadata being informative and trustworthy. If metadata is irrelevant, unavailable, or false, it may yield no gain or cause harm.

Alternative weighted decision objectives require their own derivation and justification; the supplementary discussion distinguishes such heuristics from normalized EFE.

6.4 Q3: monitor uncertainty and regulate evidence acquisition

Q3 observes a property of the inference process and changes what the system does next. Here, maximum posterior probability determines whether another observation is acquired. The monitoring rule is bounded by an explicit sensing budget. The mechanism improves accuracy when new observations are sufficiently informative, but incurs cost and can fail under miscalibrated beliefs.

Posterior entropy and maximum posterior probability describe belief concentration. Calibration requires comparison with realized outcomes. The package’s composite confidence and attention examples are illustrative control rules; their weights are not fitted psychological parameters or calibrated probabilities of correctness.

6.5 Q4: evaluate model assumptions and justify their revision

Q4 evaluates the reliability model and changes its parameters. The implemented mechanism estimates likelihood reliabilities on a separate labeled training set and freezes them before test. This is model-level adaptation under external supervision. It is not unrestricted self-modification, a learned ontology, or evidence that an agent can independently choose appropriate goals.

A general framework-revision objective may combine predictive fit, resource cost, and constraints on permissible changes. The choice of those terms requires justification; lowering an in-sample objective is not enough. Held-out evaluation and bounded update rules distinguish improved fit from overfitting or drift. Revising preferences also requires an external criterion for whether the revision is desirable.

6.6 Overlapping roles, comparative predictions, and falsifiability

The framework suggests testable contrasts: informative versus uninformative metadata, matched versus mismatched models, fixed versus adaptive sensing under comparable cost, and training-based revision versus held-out controls. A useful taxonomy should help define these contrasts without predetermining their outcomes. Evidence against a particular mechanism should remain possible. The current study gives one tractable example; broader tasks, correlated sensors, nonstationarity, and representation learning are future tests.

7 Evidence Integrity, Source Trust, and Model Vulnerability

7.1 A scoped threat model for evidence and model commitments

We consider an attacker able to change observations, reported reliability, monitoring rules, or selected model parameters in a categorical artificial agent. The assets are the fidelity of state estimates, the integrity of preferences, and the validity of model revisions. We assume the evaluation state and held-out scoring code remain trusted. This is a conceptual and synthetic threat model; it does not establish attack prevalence, impact in deployed systems, or human cognitive vulnerability.

Intervention	Example	Evidence or limitation
Observations (Q1)	Alter a sensor outcome	Can be studied as a controlled observation perturbation
Metadata (Q2)	Overstate reliability of a weak sensor	Measured adverse control in the binary task
Monitoring (Q3)	Alter a confidence threshold or sensing budget	Requires a separate monitoring-rule intervention
Model revision (Q4)	Poison training labels or permit preference drift	Requires a trusted update reference and held-out evaluation

These locations need not be exclusive. Altering a likelihood matrix may affect both direct inference and metadata-conditioned inference. There is no measured monotonic ranking of danger or detectability across Q1–Q4.

7.2 Misleading metadata as a measured synthetic failure case

The misleading-metadata condition claims high reliability for the weakest sensor and low reliability for the strongest. Held-out accuracy falls to 54.12%, compared with 84.68% under truthful reliability metadata. Log loss also increases. This demonstrates a specific misspecification failure under known ground truth. It does not establish the effectiveness of a detector, defense, or real-world influence operation.

7.3 Trust, calibration, and the limits of diagnosis

An anomaly score can reveal a discrepancy without identifying its cause. A change may be benign, and a harmful intervention may remain close to the baseline. Likewise, a well-formed probability model can encode false information. The technical diagnostic contracts are retained in Supplement 11.5; neither their thresholds nor the taxonomy establish calibrated security performance.

7.4 Bounded integrity safeguards and unresolved assurance questions

A practical system can record trusted parameter versions, verify input provenance, constrain allowed revisions, retain rollback points, and test candidate changes on held-out data. These are design proposals requiring separate evaluation. Model hashes establish byte identity against a trusted reference; they do not establish the truth or appropriateness of that reference.

Preference specification remains a normative choice. An inspectable C vector can help make that choice discussable, but its mathematical representation does not ensure alignment with human values Gabriel [2020], Amodei et al. [2016]. Future security

experiments should state attacker capabilities, clean and attacked distributions, detector calibration data, resource budgets, and false-positive/false-negative tradeoffs. No security efficacy claim is made here.

7.5 The separate claims an assurance argument must establish

Specification provenance addresses **which model was executed**. Mathematical checks address **whether the declared calculation obeys its contracts**. Held-out controls address **how that mechanism behaves in a specified environment**. None answers the normative question **whether the objective should be adopted**. A defensible assurance argument states the trusted reference, permitted interventions, evaluation distribution, adoption criterion, and known failures separately. The misleading-metadata condition illustrates why a well-formed specification can still encode bad information.

8 Interpreting Experimental Evidence and Justifying Model Revision

8.1 What the controlled experiments establish about these mechanisms

The experiment supports a narrow conclusion: using accurate source reliability can improve a misspecified inference baseline, and additional evidence can improve accuracy at additional cost. Estimating reliability from labeled training approximates the oracle condition in this task. These findings are compatible with standard Bayesian model specification and sequential evidence acquisition. They do not require that the four quadrants correspond to distinct computational or biological faculties.

Q2 and Q4 produce identical held-out classification accuracy in the retained run, while their Brier and log losses differ. In this parameter regime, fitted likelihood weights preserve the ordering of classification decisions across seeds. Equal accuracy therefore does not imply identical beliefs or perfect parameter recovery. Conversely, a sizable standardized seed-level effect largely reflects small variation in replicate means, so absolute percentage-point differences and sensing costs deserve priority.

8.2 What the four-role taxonomy contributes to model criticism

The taxonomy makes three questions explicit: what evidence is modeled, what information is available about its reliability, and which part of processing can be monitored or revised. It connects model specification to testable interventions and makes assumptions inspectable. It does not supersede established work on learning, monitoring/control, active learning, or hierarchical inference.

Changing A , B , C , or D has different meanings: observation model revision, dynamics revision, preference revision, and prior revision. These should not all be treated as the same form of optimization. Empirical fit can assess some epistemic revisions, whereas preference revision additionally needs a justified normative reference.

8.3 Meta-epistemic practice as fallible inquiry

The philosophical proposal is methodological: model specification should be open to criticism through its consequences. Inference is conditional on a representation; inquiry includes asking whether that representation excludes a relevant possibility. Gelman and Shalizi’s account of model checking and revision supplies a useful precedent for this distinction [Gelman and Shalizi \[2013\]](#). Our taxonomy does not resolve the disagreement between confirmation-based and criticism-based philosophies of Bayesian practice. It makes the distinction visible in an executable case.

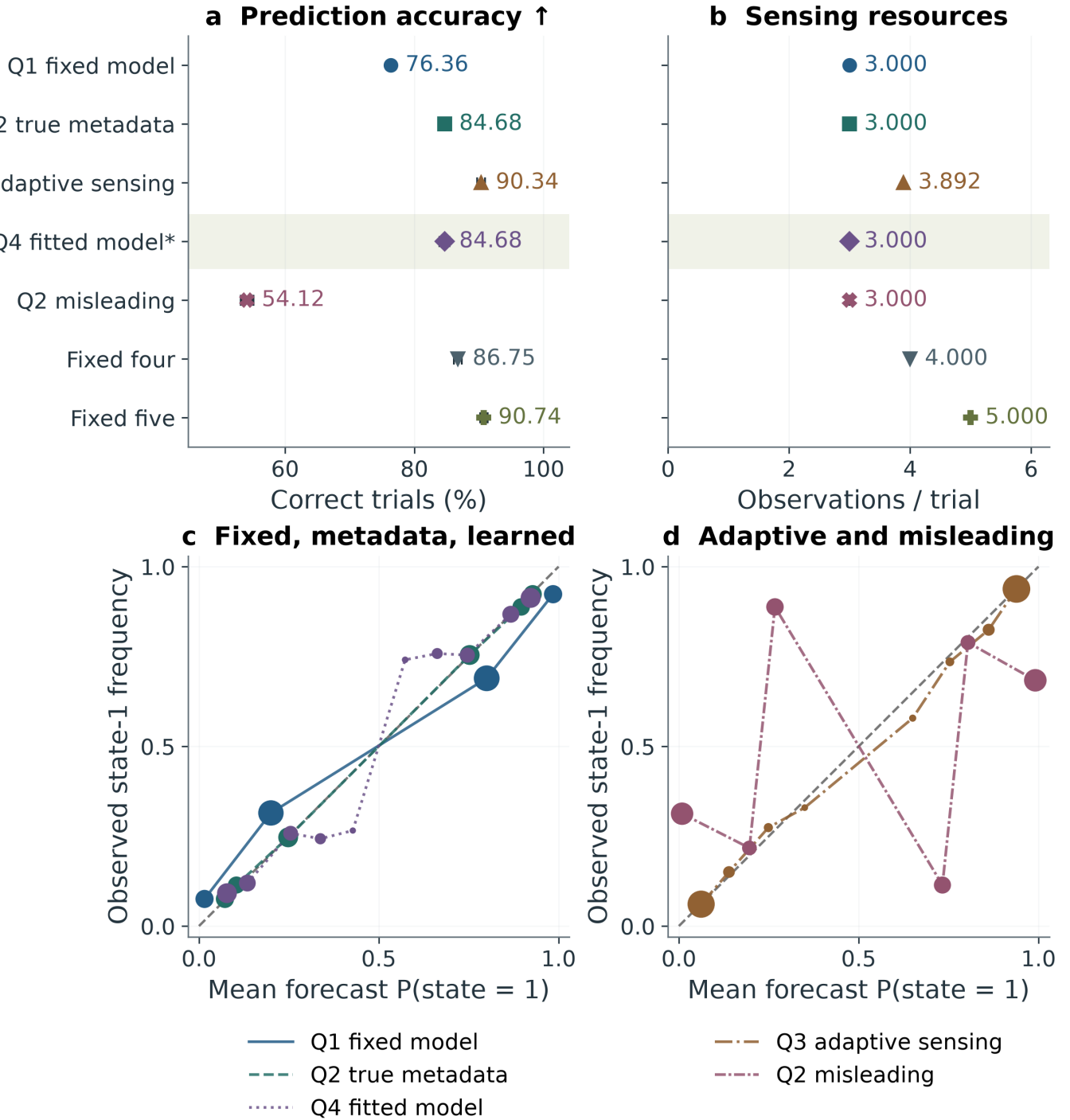
Metadata has no intrinsic privilege. Reliability information becomes evidence only through a model of its source, relevance, and trustworthiness. A posterior can be exact conditional on false metadata. GNN makes that conditioning inspectable; it does not eliminate the need for independent evidence about the source. Fallibility therefore applies to the likelihood, its reliability estimate, the monitoring rule, and the evaluator’s own reference.

8.4 Meta-pragmatic practice and the authority to revise

Preference logits specify a ranking of outcomes, not a complete ethics. An informational objective, a sensing budget, and an adoption threshold embody different judgments. Our interpretation connects the evaluation of those judgments to James’s emphasis on practical consequences [James \[1907\]](#), without equating usefulness with unconstrained truth or morality. Better prediction under one scoring rule cannot authorize a change in whose interests the system serves.

The modeler and the modeled agent must also be distinguished. Here the researcher declares the task, external labels supervise learning, and a fixed rule triggers sensing. The experiment does not show an agent independently inventing a representation or legitimating its own ends. An autonomous revision mechanism would additionally need an admissible change space, an independent evaluation reference, and a rule for accepting or rejecting the change. Those requirements are design proposals.

8.5 Comparing prediction quality, sensing resources, and calibration



Top: same strategy rows; marginal 95% seed intervals.

* Learned model also uses separate labeled training.

Bottom: 12,000 trials/strategy; 10 bins; area scales with count.

Dashed diagonal: agreement. No calibration uncertainty bands.

Figure 4: Prediction, sensing resources and descriptive calibration answer distinct questions. (a,b) Aligned strategy rows show percentage accuracy and observations per trial, with numeric means, distinct symbols and marginal 95 percent intervals over 30 paired seeds. Shading marks the fitted model's separate labeled training, excluded from the observation count. (c,d) Two views separate the pooled calibration comparisons across 12,000 held-out trials per strategy. Line style supplements color; point area scales with occupied-bin population. Points use mean forecasts and observed state frequencies in ten fixed bins, with empty bins omitted. The dashed diagonal denotes agreement. These curves have no calibration uncertainty bands.

The fixed-budget comparators make the resource tradeoff explicit. Their results should be read together with Q3’s stopping rule and average cost; a single accuracy ranking does not identify an optimal procedure. A decision about whether extra observations are worthwhile needs a stated cost function. No wall-clock, energy, or monetary performance is inferred from observation counts. This connects the computational comparison to bounded rationality while leaving application-specific valuations open [Simon \[1955\]](#).

In the retained run, fixed four-observation inference achieves 86.75% accuracy and fixed five-observation inference achieves 90.74%. Q3 achieves 90.34% at a mean 3.892 observations. The stopping rule allocates observations to less confident trials; it can therefore outperform a fixed four-observation schedule while using fewer observations on average. These are descriptive results in one design, not a claim of universal dominance or a new multiplicity-adjusted comparison. Fixed five-observation inference has lower retained Brier and logarithmic losses than Q3, so similar classification accuracy does not mean that early stopping preserves all probabilistic information.

Probabilistic quality is likewise multidimensional. Accuracy discards the strength of a belief, proper losses penalize probability errors, and the reliability diagram displays empirical agreement with predicted probabilities [Gneiting and Raftery \[2007\]](#), [Dawid \[1982\]](#). The diagram is descriptive for this task and binning scheme; it is not a guarantee under distribution shift.

Bröcker and Smith show why bin populations, forecast distributions and sampling variation complicate the visual reading of reliability diagrams [Bröcker and Smith \[2007\]](#). Figure 4 therefore separates comparisons and exposes bin occupancy through point area. Its pooled lines remain descriptive. We do not implement their consistency-resampling procedure, and a visually smooth curve or a small distance from the diagonal is not a statistical calibration test.

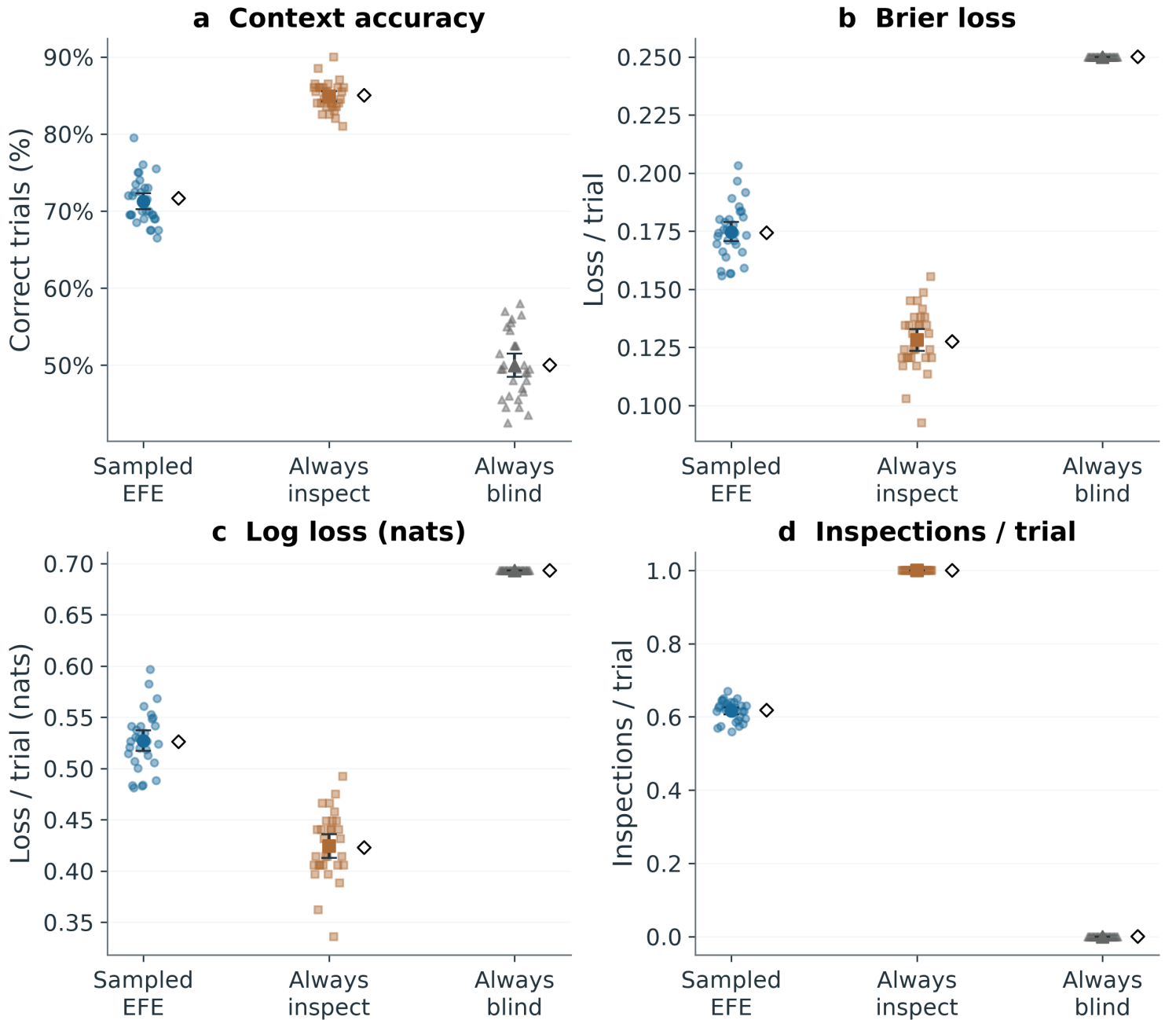
8.6 Stopping inquiry requires an explicit resource-sensitive judgment

The stopping rule expresses a procedural commitment about when present confidence warrants ending observation. Extra evidence may improve prediction while demanding resources; a resource saving does not by itself establish an acceptable error cost. Simon’s bounded rationality and resource-rational analysis motivate evaluating this tradeoff rather than hiding it in a single performance score [Simon \[1955\]](#), [Lieder and Griffiths \[2020\]](#).

The detailed paired contrasts and threshold sweep remain in Supplement 11.4, Figures 9 and 10. They expose analytical choices without selecting and validating a new policy on the same trials.

8.7 Action selection reveals the preferences and assumptions of a model

Figure 5 compares EFE policy sampling with fixed observation strategies. The sampled agent sometimes accepts a blind observation because its objective combines preferred outcomes with information. Always inspecting obtains more informative evidence, while always remaining blind supplies an explicit limiting case. These comparisons explain what the selected mechanism does under its stated commitments; they do not establish that accuracy alone is the appropriate end.



Small symbols: seed means. Capped bars: marginal 95% intervals.
Open diamonds: exact expectations. EFE = expected free energy.

Figure 5: Sampled action policies compared with exact expectations under the declared world. Small strategy-specific symbols show the 30 paired seed means, larger symbols show their averages, and capped dark bars give marginal 95 percent seed-bootstrap intervals. Open diamonds are exact expectations displaced slightly to the right for visibility. (a) Context accuracy favors larger values; (b,c) proper losses favor smaller values; (d) inspection counts expose informative-observation use. All three strategies share 6,000 generated world trials, yielding 18,000 scored rows rather than independent replications. EFE means expected free energy; its preference-sensitive action objective is distinct from maximizing accuracy.

The exact policy surface in Supplement 11.4, Figure 11, shows how changing information quality or the preference for the blind outcome changes action probabilities. This is a designed consequence of the model, not the discovery of an independently justified preference. The distinction links an executable policy to the philosophical question of who may authorize its commitments.

8.8 Exact inference and confident error in a misspecified world

Figure 6 separates the agent's model from the world's likelihood. At actual reliability 0.50, inspection is uninformative. The agent nevertheless reports inspection posteriors of 0.15 or 0.85 because it retains its assumed reliability of 0.85. Its EFE action

probabilities also remain fixed. Always inspecting then has exact Brier loss 0.3725, worse than the blind baseline’s 0.25, although both classify at chance. This is a proper-score failure hidden by equal accuracy. The case establishes a consequence of the specified mismatch, not a universal argument against information-seeking action.

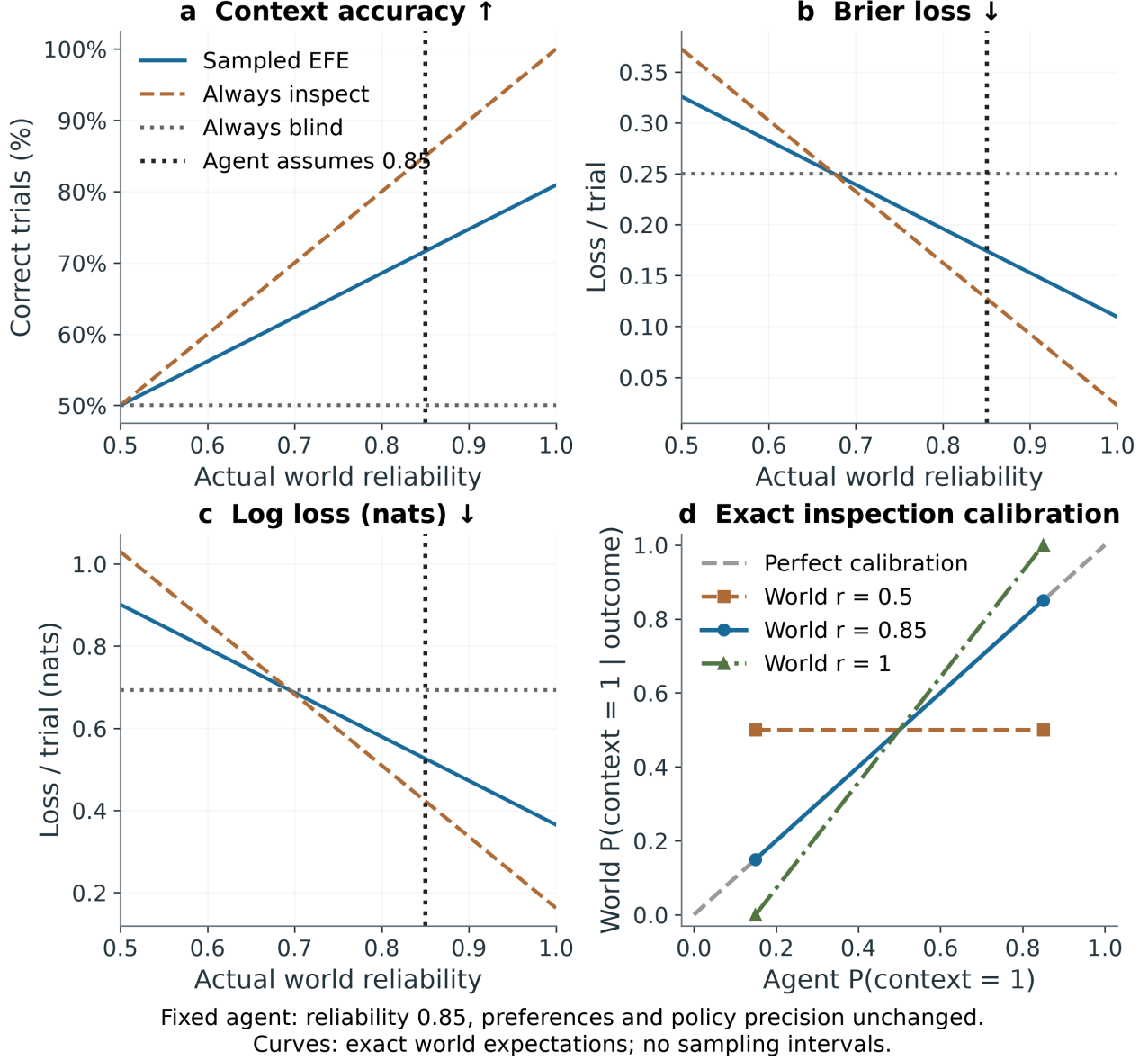


Figure 6: Correct computation can produce inaccurate predictions under an incorrect world model. (a–c) The agent holds its assumed reliability at 0.85, with preferences and policy precision fixed, while actual reliability varies. Line styles identify the sampled EFE policy and fixed-action baselines; arrows indicate favorable score directions. The vertical reference marks the matched likelihood. (d) Distinct symbols and line styles compare inspection forecasts with exact world conditional frequencies at three actual reliabilities; the diagonal denotes calibration. Every curve is an exact world expectation, not an empirical estimate or uncertainty interval. No automatic model repair is implemented.

The practical response requires evidence about reliability, a permitted update, and fresh evaluation of its consequences. Re-computing an exact posterior under the same incorrect likelihood does not provide that evidence. This is the meta-pragmatic distinction in executable form: action selection follows declared commitments, while criticism asks what would warrant retaining or changing them. The analysis holds the agent fixed to reveal the failure; it does not implement autonomous detection or repair of misspecification.

8.9 Limitations of the synthetic design and philosophical interpretation

The task has two states, conditionally independent sensors, fixed reliabilities, and no changing state within a trial. Labels are available for training Q4, and true reliability is supplied to Q2/Q3. Q3 receives a larger sensing budget. Those assumptions

favor the tested mechanisms; correlated observations, incorrect metadata, changing reliability, and costly acquisition can alter the result.

The original sensor experiment evaluates state estimation and confidence-triggered sensing; its gains are not attributed to EFE. The separate action task does execute EFE-based policy sampling and Bayesian context inference, but it is a one-step categorical control example with known dynamics and likelihood. Neither task establishes long-horizon, continuous, hierarchical or representation-learning performance. The package’s narrative framework/modeler classes and schematic FEP visualizations remain educational demonstrations.

Confidence intervals concern this simulation design. Bootstrap seed resampling approximates variability over independent generated datasets, not uncertainty about all possible environments. Sign-flip tests require a symmetric/exchangeable null; they are supplementary to retained estimates and model assumptions. The three primary tests use Holm correction, while the displayed 95% intervals are marginal. No retrospective power calculation or fabricated regression fit is reported.

The manuscript is a working research document, without claimed peer review or independent replication. Bibliographic corrections are documented separately; bibliography links establish source identity, not independent confirmation of every interpretation. Synthetic failures do not support clinical or deployment guidance.

8.10 From controlled synthetic examples to independent empirical tests

Useful next tests would compare equal sensing budgets, include model-free and model-based decision baselines, introduce correlated and shifting sensor noise, vary training-label availability, and evaluate metadata provenance or corruption detectors on held-out attacks. Representation-changing Q4 mechanisms require complexity penalties and external evaluation. A biological study would need measured behavior, confidence reports, a fitted model, and appropriate metacognitive measures [Fleming and Lau \[2014\]](#), rather than mapping the quadrant labels directly onto neural structures.

9 Toward Accountable Active Inference Through Public Criticism

Active inference makes observation likelihoods, dynamics, preferences, and priors explicit modeling objects. We interpret decisions about these objects as meta-epistemic and meta-pragmatic activities and organize examples with a data/metadata by cognition/meta-cognition taxonomy. The taxonomy is useful when it clarifies a mechanism, a comparison, and a possible failure, rather than when it merely assigns a higher-order label.

The implementation exposes explicit probability conventions, categorical and virtual-evidence inference, normalized EFE decompositions, exact variational bounds and reproducible experiment artifacts. In the specified binary sensor task, reliability-aware inference improves accuracy over a homogeneous-reliability baseline, adaptive sensing improves accuracy with additional observations, and supervised reliability fitting approximates an oracle model. Misleading metadata causes a substantial failure. These outcomes delimit the claim: better-specified or better-informed models can help, but monitoring and revision require resources, evidence and constraints.

The next research step is to test those mechanisms under richer environments and matched budgets. The present revision contributes executable GNN sources, an actual upstream parser import, fixed-budget comparators, and separate views of forecast quality and sensing resources. It supplies an inspectable starting point and an audit trail, without claiming universal performance benefits, biological validation, or security guarantees.

The separate GNN action task directly connects the declared state dynamics and likelihood to EFE policy sampling and Bayesian context inference. Its exact preference/information crossover and retained baseline comparisons show why an active inference policy need not maximize classification accuracy alone. This is a bounded demonstration of an executed mechanism; known model parameters, one-step policies and researcher-specified preferences remain substantive limitations.

The exact mismatch analysis adds a further requirement: distinguish the model’s expected consequences from consequences in a separately specified world. Fixed beliefs can yield mathematically correct EFE choices and poor probabilistic predictions when the likelihood is wrong. The retained calibration cells and independent score identities make that limitation inspectable.

Active inference serves as a meta-pragmatic framework when a declared model, executable action procedure, external evaluation reference, and accountable revision process are brought together. The central practical commitment is to expose what can be challenged: the representation of evidence, the authority of metadata, the cost of acquisition, and the criterion for model revision. The terminology alone does not provide that discipline.

10 Acknowledgments, Author Identity, and Research Integrity

This work builds on the active inference, Bayesian modeling, and metacognition literature cited throughout. Citation does not imply endorsement by the cited researchers.

10.1 Author identity and the versioned preprint archive

The author’s affiliations are Active Inference Institute and FractiAI. The author’s ORCID is [0000-0001-6232-9096](#). The version 1.0.0 manuscript and reproducibility archive are identified by DOI [10.5281/zenodo.23397453](#). This is a preprint; archival release does not imply peer review. The repository’s changelog distinguishes the initial public release from earlier private development revisions.

10.2 Synthetic intelligence assistance and human accountability

Synthetic intelligence, including generative artificial intelligence (AI) assistance through OpenAI Codex, supported prose drafting and revision, code development and refactoring, source-metadata checks, programmatic visualization and rendering workflows. The numerical evidence comes from retained seeded simulations and exact categorical calculations, rather than language-model responses. Source identities and access limits, executable tests and artifact hashes are documented in the accompanying repository. These checks support inspection and reproduction; they do not constitute independent scientific review. The human author retains responsibility for the claims, interpretations, citations and final approval of the manuscript. AI tools are not credited as authors, consistent with COPE’s distinction between assistance and accountable authorship [COPE Council \[2024\]](#).

10.3 Synthetic data, retained artifacts, and research declarations

The numerical studies use synthetic data only and involve no human or animal participants. Funding and competing-interest declarations have not been supplied for this draft; the author should complete the declarations required by the eventual submission venue.

The manuscript, implementation, generated data and figure sources are retained with configuration and source hashes in the accompanying [source repository](#). The versioned archive includes the manuscript, source and computational evidence; release checksums bind its downloadable files. The repository carries an Apache-2.0 license. Historical development snapshots remain identifiable in Git rather than being presented as prior public releases.

The cover is a scientific graphical abstract generated programmatically from retained synthetic estimates and exact GNN/EFE evaluations. Its source data, generator and asset digests are recorded with the manuscript assets. Earlier conceptual artwork and its prompt remain as historical assets and are not the selected cover. The model-revision diagram is an authored conceptual illustration; the empirical figures and policy surfaces use retained synthetic trials or exact model evaluations. The upstream GNN and template implementations are separately attributed and are consumed without modifying their tracked source.

11 Technical Supplement: Formal Models and Statistical Protocols

This supplement retains the mathematical and experimental detail behind the main argument. The model atlas makes executable assumptions inspectable; the full protocol preserves parameters, inferential units, controls and qualifications. Supplement [12](#) consolidates notation for the main paper and this companion.

11.1 Imported GNN tensors behind the experimental model diagrams

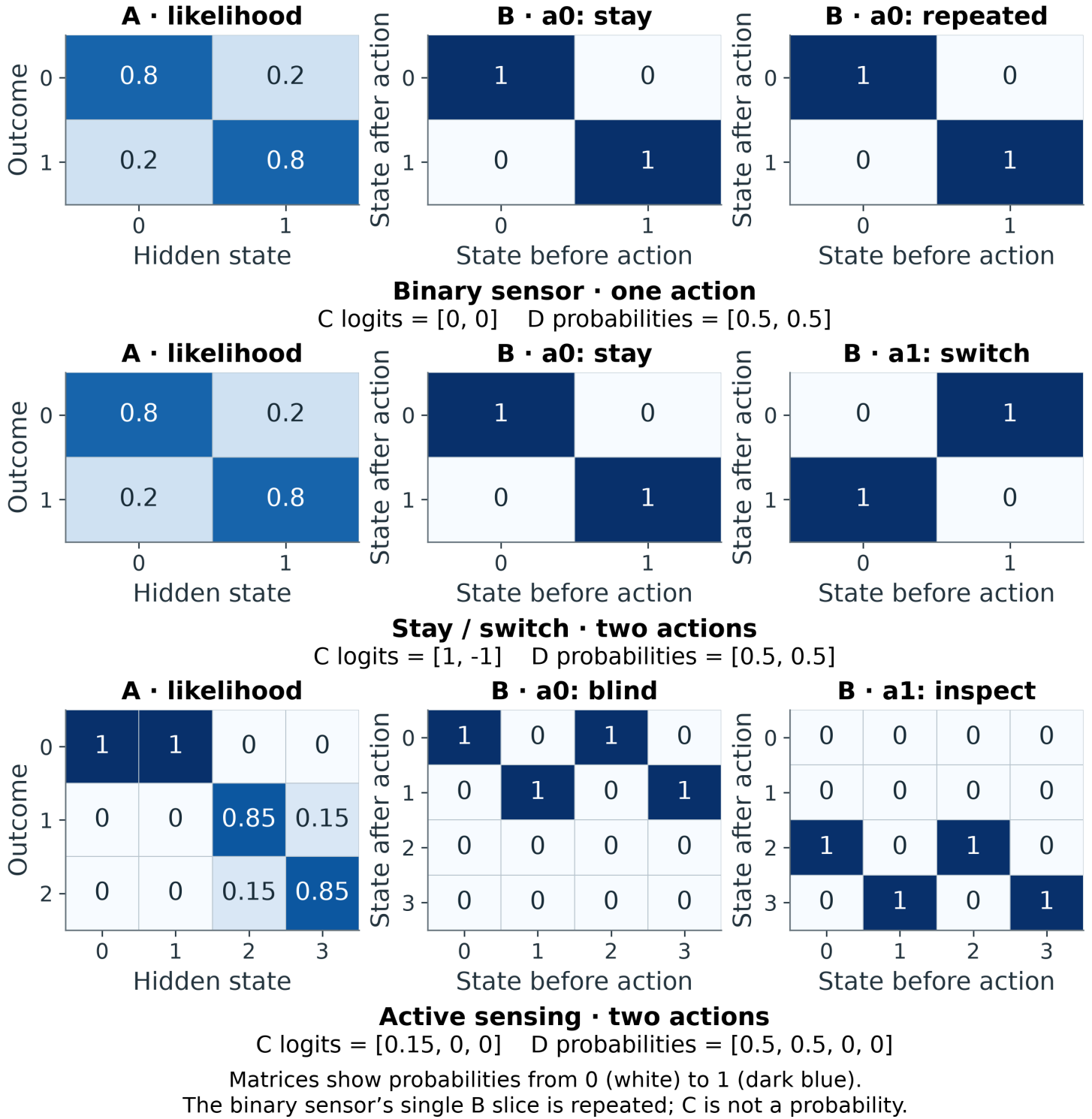


Figure 7: Actual imported GNN tensors make the model declarations inspectable. Each group prints likelihood A and action-specific transition B probabilities, with cell boundaries and all zero entries retained. A has outcome rows and hidden-state columns; B has after-action rows and before-action columns. Actions are named stay/switch or blind/inspect; the binary sensor's single B slice is repeated for layout. White means zero support and dark blue means probability one. Signed preference logits C and normalized initial beliefs D appear below each group. The stay/switch model is an analytical software control, not an additional experimental study.

11.2 Categorical probability contracts and exact Bayesian inference

11.2.1 The discrete generative model and its tensor conventions

Let s index hidden states, o observations, and a actions. The implementation uses $A_{o,s} = p(o \mid s)$, $B_{s',s,a} = p(s' \mid s, a)$, finite outcome preference logits C_o , and a normalized initial prior D_s . Columns of A and of each action slice of B sum to one. A two-dimensional B is promoted to a single-action tensor. Arrays with incompatible dimensions, empty axes, negative probabilities, nonfinite values, or invalid normalization are rejected.

The normalized outcome preference is $p_C(o) = \exp(C_o) / \sum_j \exp(C_j)$. Logits are not probabilities; adding a constant to all entries leaves the normalized preference unchanged. Model validation copies inputs so that later mutation of caller-owned arrays does not silently change an initialized model. Public model arrays remain inspectable and mutable; callers must validate deliberate revisions.

11.2.2 The supported GNN profile and executable import boundary

Canonical GNN model files live under `src/core/gnn_models/`. The binary-sensor specification declares static identity dynamics; the stay-switch agent declares two action-conditioned transition slices. Both declare the same axis conventions used by the runtime. Each numerical intervention is materialized as GNN Markdown, parsed by the actual upstream `MarkdownGNNParser`, checked against this project’s supported profile, and converted into the matrices consumed by inference.

The adapter rejects missing or duplicate sections, parser warnings, unsupported flags and model families, contradictory dimensions, malformed tensors, invalid probabilities, and unsupported dependency graphs. It verifies the upstream Git origin, pinned revision, and tracked source state before importing the parser. An absent checkout does not select a substitute parser or an undeclared model. Tensor changes are tested through changes in executed beliefs and policy choices, not merely through successful parsing.

This boundary makes the specification executable without confusing syntax validity with scientific adequacy. The experiment retains the exact materialized sensor specifications and their digests. Random-model identity checks and the reliability sweep also pass through GNN import. Direct tensor constructors remain available as a low-level numerical API and as independent test oracles.

11.2.3 Bayesian state inference with hard and virtual evidence

For a categorical observation, Bayes’ rule gives:

$$q(s \mid o) = \frac{A_{o,s} D_s}{\sum_j A_{o,j} D_j}. \quad (1)$$

For a normalized nonnegative virtual-evidence vector $e(o)$, the implementation uses $L_e(s) = \sum_o e(o) A_{o,s}$ and replaces $A_{o,s}$ with $L_e(s)$ in Equation (1). This convention models evidence about possible outcomes. It differs from a Jeffrey update, which mixes already-conditioned posteriors. A uniform evidence vector is uninformative because each likelihood column sums to one. Zero-probability states retain zero mass; evidence impossible under the model and prior raises an error instead of being repaired by clipping.

11.2.4 Exact variational free energy and the evidence bound

Given an explicit joint distribution and a normalized candidate posterior q , variational free energy is:

$$F(q, o) = \sum_s q(s) [\log q(s) - \log p(o, s)] = -\log p(o) + D_{\text{KL}}[q(s) \parallel p(s \mid o)]. \quad (2)$$

Equation (2) bounds negative log evidence and reaches the bound at the exact posterior. Candidate beliefs assigning positive mass outside posterior support have infinite free energy. The exact implementation is `calculate_variational_free_energy`; the older model-free `FreeEnergyPrinciple.calculate_free_energy` path remains an explicitly labeled schematic score for compatibility. Its trajectories and state-boundary diagrams do not establish a variational bound, conditional independence, or the physical claims of the FEP. Supplying a generative model to that class selects the exact calculation.

11.2.5 Normalized expected free energy and categorical policy selection

For each future policy step, propagate $q_\tau = B(a_\tau) q_{\tau-1}$ and predict $q_o = A q_\tau$. With an exact conditional update, state information gain is $I(s; o) = H(q_o) - \sum_s q_\tau(s) H(A_{\cdot,s})$. Our normalized planning score is:

$$G_\tau = -I(s; o \mid \pi) - \sum_o q_o(o) \log p_C(o) = D_{\text{KL}}[q_o \parallel p_C] + \sum_s q_\tau(s) H(A_{\cdot,s}). \quad (3)$$

The equivalence in Equation (3) follows by expanding mutual information and KL divergence. It is exact for this finite, known-likelihood predictive model. It does not include uncertainty about parameters, posterior approximation error, or observation-contingent replanning.

For compatibility, the returned `total` uses the raw-logit score $-I - \mathbb{E}[C]$. `normalized_total` adds `logsumexp(C)` per step and equals risk plus ambiguity. Normalized values are evaluated directly from stable log preferences to avoid cancellation under large constant logit shifts. Policy rankings agree mathematically for a fixed model and horizon convention. Mean aggregation is the legacy default; sum aggregation implements additive finite-horizon costs. Comparing different horizons or different preference models requires specifying this convention and normalization explicitly.

Policy selection accepts current posterior beliefs, with D used when beliefs are omitted. It returns an argmin and a policy distribution proportional to $\exp(-\gamma G)$, where $\gamma > 0$ is precision. Increasing precision concentrates this distribution; it does not change an exact argmin. Equal scores retain a uniform distribution. All entropy and KL quantities use natural logarithms and are measured in nats.

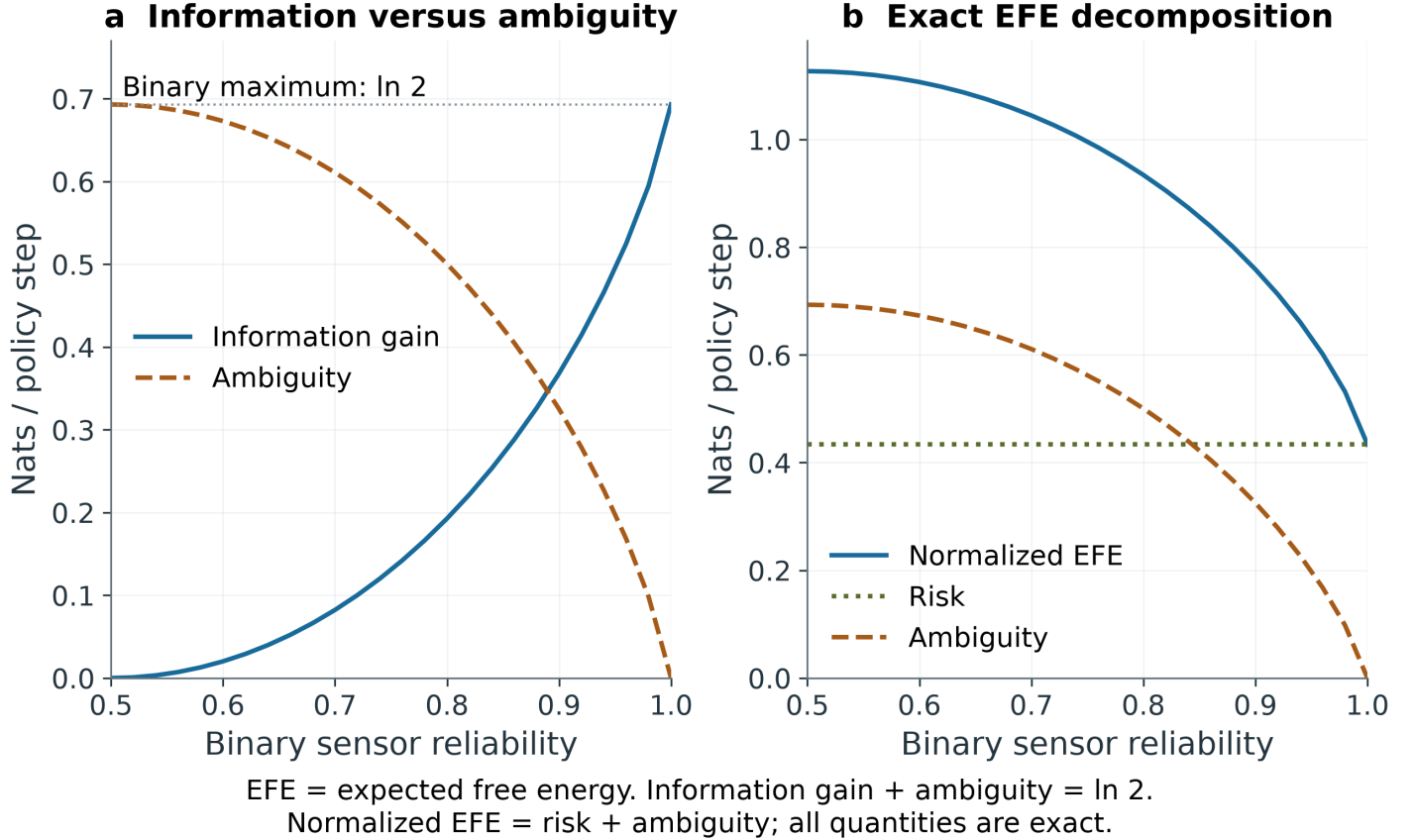


Figure 8: An exact one-step decomposition separates information, ambiguity and outcome risk. For a binary symmetric likelihood, uniform state beliefs, identity transitions and $C=(1,-1)$, (a) information gain rises toward the binary maximum $\ln 2$ while ambiguity falls, and their sum stays $\ln 2$. (b) Uniform predicted outcomes keep risk constant, so normalized expected free energy equals risk plus ambiguity throughout. Axis units are nats per policy step; distinct line styles identify the terms. The figure evaluates declared mathematics without fitted effects, sampling uncertainty or evidence that an agent’s assumptions fit its world.

11.3 Complete synthetic experimental design and statistical protocol

11.3.1 Study type and inferential scope

This is a conceptual framework with a synthetic computational demonstration. The taxonomy was not derived from empirical clustering and is not tested as a biological partition. The experiment tests concrete mechanisms inspired by its quadrants. The mechanisms are implemented in `src/analysis/experiments.py`; the narrative quadrant class supplies definitions and examples, rather than four independently validated algorithms.

The design below is fixed in code and recorded in each output manifest. It is not a preregistered study. Historical numerical claims in earlier manuscript versions lacked a corresponding experiment and are superseded by the retained results reported here. No participant data, clinical outcomes, or deployment incidents are analyzed.

11.3.2 The binary sensor task and its resource-matched controls

Each test trial independently draws a binary hidden state from a uniform prior. Five conditionally independent binary sensors agree with that state with probabilities (0.55, 0.65, 0.85, 0.80, 0.80). Every strategy sees the same pre-generated test states and observations within a seed. The first three sensors are the common initial evidence. The state remains fixed within a trial and resets independently between trials.

- **Q1 fixed:** inference assumes reliability 0.80 for each of the first three sensors.
- **Q2 metadata:** inference uses the true, sensor-specific reliabilities for the same three observations. This is an oracle-metadata condition.
- **Q3 adaptive:** starts from Q2, then acquires a fourth and, if needed, fifth observation while the maximum posterior probability is below 0.90. Its cost is the actual observation count, with a maximum of five.
- **Q4 learned:** estimates the first three reliabilities from 200 separate labeled training trials per seed, using a Beta(1,1) prior. Parameters are frozen before held-out evaluation. It uses exactly three test observations.
- **Misleading metadata:** uses claimed reliabilities (0.95, 0.85, 0.55) for the first three sensors. This prespecified adverse control reverses the ordering of sensor quality.
- **Fixed four / fixed five:** use true reliability with exactly four or five observations on every trial. These added descriptive comparators separate a fixed sensing budget from the confidence-triggered stopping rule. They are additions to this revision, not preregistered tests.

Q3 changes available evidence, whereas Q4 changes fitted likelihood parameters. They are not matched to Q1 on every resource. We report these differences rather than attributing all gains to a generic meta-level capacity. The confidence score controlling Q3 is maximum posterior probability; it is not the composite heuristic returned by `MetaCognitiveSystem`.

11.3.3 An action task driven by expected free energy

A separate exploratory task directly executes active inference action selection. Four hidden states encode binary context crossed with observation mode: two blind-mode states and two informative-mode states. Two actions preserve context while setting the mode. The blind action produces a null outcome; inspection produces one of two context-indicating outcomes with reliability $r = 0.85$. The declared likelihood is

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & r & 1-r \\ 0 & 0 & 1-r & r \end{pmatrix}, \quad C = (c, 0, 0), \quad D = (1/2, 1/2, 0, 0), \quad c = 0.15. \quad (4)$$

Each transition slice deterministically maps the current context to the action’s mode. Every model is materialized as GNN and parsed through the actual upstream reference parser before the local categorical runtime consumes its tensors. The two one-step policies are scored by normalized EFE. Actions are sampled from the actual returned policy posterior at precision 4, rather than replacing it with an unrelated confidence rule.

For this uniform-context design, inspection supplies information $I = \log 2 - h(r)$, where h is binary entropy in nats. The normalized score difference is $G_{\text{inspect}} - G_{\text{blind}} = c - I$. Thus, increasing the preference for the null outcome can reverse the minimum-EFE action even when information quality is fixed. The sampled policy has inspection probability $[1 + \exp(4(c - I))]^{-1}$; finite precision retains both actions. This is a designed tradeoff, not an independently learned preference or a claim that the sampled policy minimizes expected score without regularization.

Thirty independent seeds, offset by 10,000 from the first task, generate 200 context/outcome trials each. The world draws initial states from the imported D , next states from the imported action slice of B , and outcomes from the imported A . Separate world and action random streams prevent action sampling from altering the world draws. Sampled EFE, always inspect, and always blind share each generated context and potential informative outcome. Context labels score predictions only; no parameter is fitted. Both actions produce a categorical observation, while the inspection count records whether that observation is informative. It is not an energy or wall-clock measurement.

The study retains 18,000 strategy-trial rows, 90 seed metrics, marginal 95% seed-bootstrap intervals, paired differences against both fixed baselines, and exact expectations under the declared model. Accuracy, Brier loss, log loss, inspection count, preferred-outcome logit, selected normalized EFE and posterior entropy are reported. No new significance family is introduced. An exact 11-by-17 reliability/preference grid separately evaluates action scores and probabilities; the grid is not a set of independent empirical replicates.

11.3.4 Separating the world from the agent’s model

The matched action task deliberately uses the same world and agent parameters. To expose the limitation of that match, an additional exact analysis fixes the agent’s assumed reliability at $\rho = 0.85$, its preferences at $c = 0.15$, and its policy precision at 4, while varying actual world reliability r over 21 values from 0.50 to 1.00. Each world and the fixed agent are imported from

their own content-addressed GNN specification. The policy remains the agent’s returned EFE distribution; evaluation uses world probabilities. No parameter is fitted and no policy is selected from this grid.

Let a denote an action, s' the next state, and $y(s')$ its context, encoded by state-index parity. For world matrices carrying subscript w , the exact joint distribution is

$$p_w(s', o | a) = A_w(o, s') \sum_s B_w(s', s, a) D_w(s). \quad (5)$$

The agent independently computes $\hat{p}_a(o) = q_{\text{agent}}(y = 1 | o, a)$. For any binary scoring loss ℓ , its exact expected value under strategy π is

$$R_\ell(\pi; w, \text{agent}) = \sum_{a, s', o} \pi(a) p_w(s', o | a) \ell(\hat{p}_a(o), y(s')). \quad (6)$$

Zero-mass events contribute nothing. Positive-probability world observations impossible under the agent fail explicitly; zero forecast probability for a possible context likewise fails rather than being clipped to an arbitrary finite logarithmic loss. The present grid avoids these singularities because ρ is interior to the unit interval. Accuracy uses the same deterministic tie rule as the sampled study.

For inspection, independent closed-form checks are

$$R_{\text{Brier}} = r(1 - \rho)^2 + (1 - r)\rho^2, \quad R_{\text{log}} = -r \log \rho - (1 - r) \log(1 - \rho). \quad (7)$$

The blind baseline has accuracy 1/2, Brier loss 1/4, and log loss $\log 2$. The EFE strategy mixes its two action-specific scores using the fixed agent policy. Calibration cells retain each action, outcome, world probability mass, agent forecast and world conditional context frequency. These are exact conditional probabilities, not binned sample estimates. The 63 strategy-by-world rows are deterministic evaluations, with no bootstrap intervals, p-values or claim of 63 independent replications. Reusable world sampling and exact scoring are separate modules from the study driver.

11.3.5 Independent seeds, random streams, and scored outcomes

Thirty independent seeds (0–29) generate 400 held-out trials each. A `SeedSequence` spawns distinct training and test streams. Labels from test trials are used only for scoring. The training and held-out streams are not used to tune the threshold, prior, or hypotheses. Equal posterior probabilities select state zero deterministically.

The primary outcome is held-out accuracy. Secondary descriptive outcomes are mean Brier loss, logarithmic loss in nats, and observations per test trial. Lower loss is better. Accuracy alone may conceal confidence errors, and a high posterior is not evidence of calibration under misspecification. The added reliability diagram uses ten fixed bins over the unit interval, pooling held-out trials descriptively; it supplies no trial-independent confidence intervals.

The three directional predictions concern (H1) Q2 versus Q1 accuracy, (H2) Q3 versus Q2 accuracy with additional sensing, and (H3) Q4 versus Q1 accuracy after separate training. The comparisons are operationalizations of these mechanisms, not claims that every quadrant extension improves accuracy. Two-sided tests retain the possibility of harm.

11.3.6 Paired seed-level uncertainty, tests, and multiplicity control

The inferential unit is a seed-level mean, yielding 30 paired replicates per comparison; the 12,000 test trials are not treated as 12,000 independent agent replicates. The common-random-number design reduces comparison noise. We report mean paired differences with 95% percentile-bootstrap intervals from 5,000 resamples of seeds, using bootstrap seed 42 [Efron and Tibshirani \[1993\]](#). Intervals are marginal, not simultaneous family-wise intervals.

Primary test evidence is a two-sided paired sign-flip test using 10,000 Monte Carlo resamples (seed 2718) with a plus-one correction. Its validity requires exchangeability of signs under the paired null; it is not an exact finite-sample test for arbitrary differences. Holm correction applies to the family of three primary comparisons [Holm \[1979\]](#). Paired t-tests and standardized within-pair differences $d_z = \bar{\Delta}/s_{\Delta}$ are supplementary. Effect-size conventions are explicit; large standardized differences across low-noise seed averages do not imply large population effects in human cognition. The new fixed-budget comparisons are descriptive and are not silently added to the original three-test family.

The pipeline also checks 200 two-step policy evaluations across 50 randomly generated models against the risk-plus-ambiguity identity, and checks the VFE evidence decomposition. These check the implementation under specified numerical contracts, not the truth of the FEP as a physical theory. Deterministic and impossible-evidence cases are included in regression tests.

11.3.7 Retained trials, fitted specifications, and computational evidence

An exploratory extension recomputes paired seed differences for six contrasts across accuracy, Brier loss, log loss, and sensing cost. It uses 5,000 seed-bootstrap resamples with seed 314159 for marginal 95% intervals. Secondary binomial sign tests count positive and negative nonzero differences, omit ties, and return one when all differences tie. They concern sign balance, not a zero mean, and their unadjusted values are not used for new rejection claims. The original three-test family remains the primary analysis.

The extension also evaluates eleven fixed stopping thresholds from 0.50 to 1.00 in steps of 0.05 on the same five sensor outcomes per test trial. No threshold is selected. This assesses sensitivity rather than providing independent policy validation. Probability-calibration tables retain occupied-bin counts and represented seeds at 5, 10, and 20 equal-width bins, with probability one assigned to the final bin. They are descriptive pooled summaries without inferential intervals.

`results/trials.csv` retains test states, sensor outcomes, strategy probabilities, and Q3 sensing counts. `results/replicates.csv` retains all seed-level metrics. `results/summary.json` contains intervals, tests, design parameters and limitations; `results/efe_sensitivity.csv` contains the exact curves. `results/manifest.json` hashes source files, configuration, environment versions and output artifacts. The Git revision is the base revision; source digests identify the executed working-tree content.

`results/gnn/` retains the imported sensor realizations, including fitted training parameters. Their content hashes bind the reported probabilities to executable specifications. For each seed, the learned condition consumes 200 labeled training trials involving three sensors; this training resource is separate from its three held-out observations per trial. The oracle conditions receive true reliability as privileged information. The cost figure measures observations, not an unmeasured monetary or energy cost.

The retained sensor outcomes are displayed in the main paper, Figure 3.

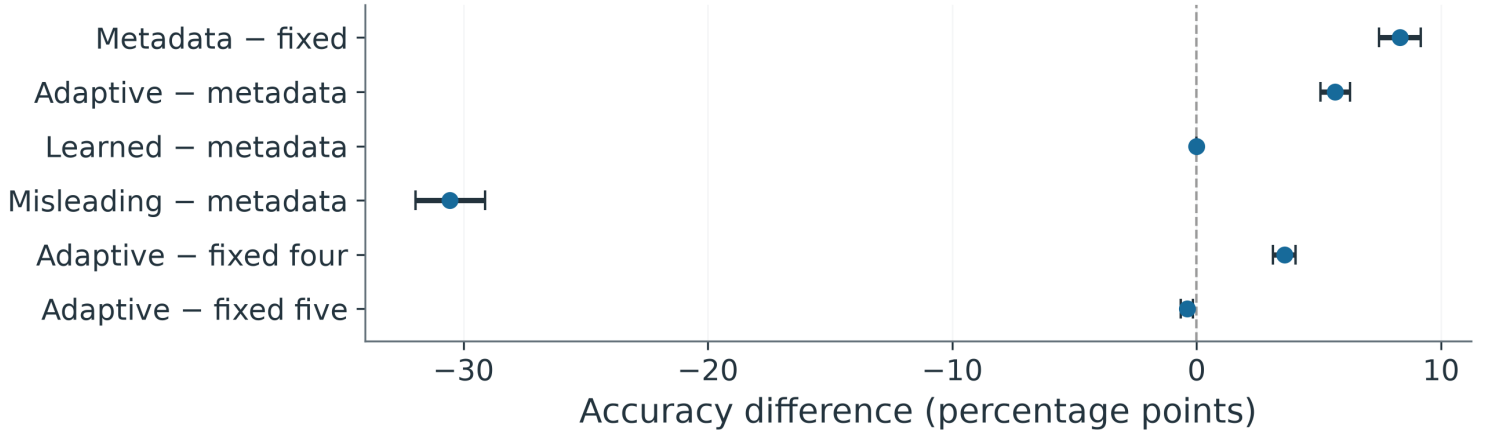
11.4 Paired contrasts, stopping rules, and policy surfaces

11.4.1 Paired predictive contrasts and the confidence-based stopping decision

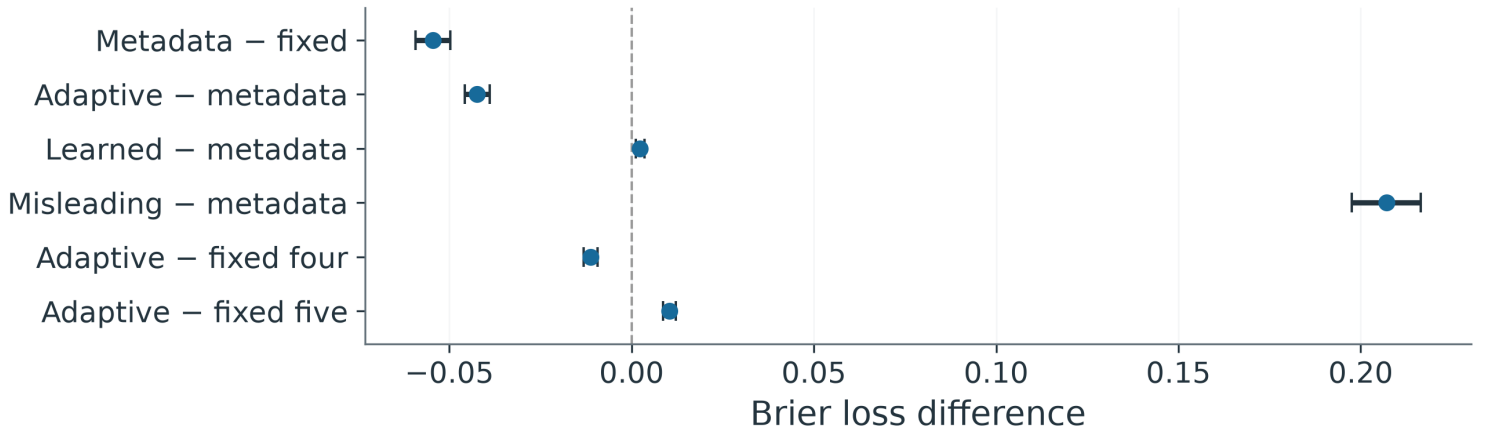
Figure 9 evaluates the same matched seeds across classification accuracy and proper losses. The direction of an accuracy gain and the direction of a loss improvement are opposite; presenting both makes this distinction visible. In particular, comparing learned reliability with oracle metadata exposes probabilistic differences that equal classification accuracy conceals. These secondary estimates have marginal intervals and are exploratory. Their unadjusted binomial sign tests concern the balance of positive and negative nonzero seed differences; they are not tests of the mean effect and do not join the original Holm family.

Paired estimates across accuracy and proper losses

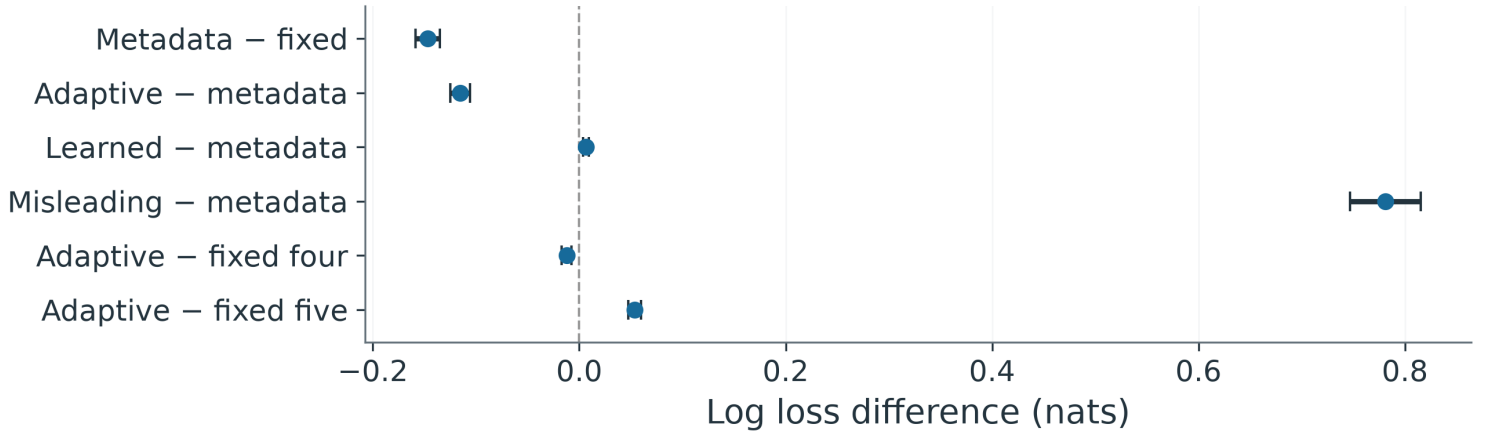
a Accuracy: right favors first



b Brier loss: left favors first



c Log loss: left favors first



Points: paired seed means; bars: marginal 95% intervals.
Accuracy favors positive differences; losses favor negative differences.

Figure 9: Paired differences preserve the common experimental units. Each row is the first named strategy minus the second on the same 30 seeds; points estimate the mean paired difference and capped dark bars show marginal 95 percent percentile-bootstrap intervals. (a) Positive accuracy differences favor the first strategy; (b,c) negative Brier or logarithmic-loss differences favor it. The dashed zero reference denotes equal scores. Accuracy uses percentage points and log loss uses nats; panel scales differ. These six exploratory comparisons do not have simultaneous interval coverage and do not expand the prespecified primary Holm testing family.

Relative to fixed four-observation inference, the adaptive rule has a mean paired accuracy difference of 3.592 percentage points, with a marginal 95% interval of [3.117, 4.042]. Relative to fixed five-observation inference, its difference is -0.400 percentage

points, with interval $[-0.650, -0.158]$. These estimates describe the tradeoff in this simulation. They do not establish equivalence to fixed-five inference; an equivalence claim would require a justified margin and an appropriate design.

The stopping threshold expresses a procedural commitment about when current confidence warrants ending observation. Figure 10 varies that commitment while holding each seed’s generated states and five sensor outcomes fixed. Confidence can be useful for allocating evidence in a correctly specified task, but the misleading-metadata condition shows why confidence alone cannot justify the model that produces it. A model revision practice therefore needs checks directed at the likelihood and data source as well as checks directed at posterior uncertainty.

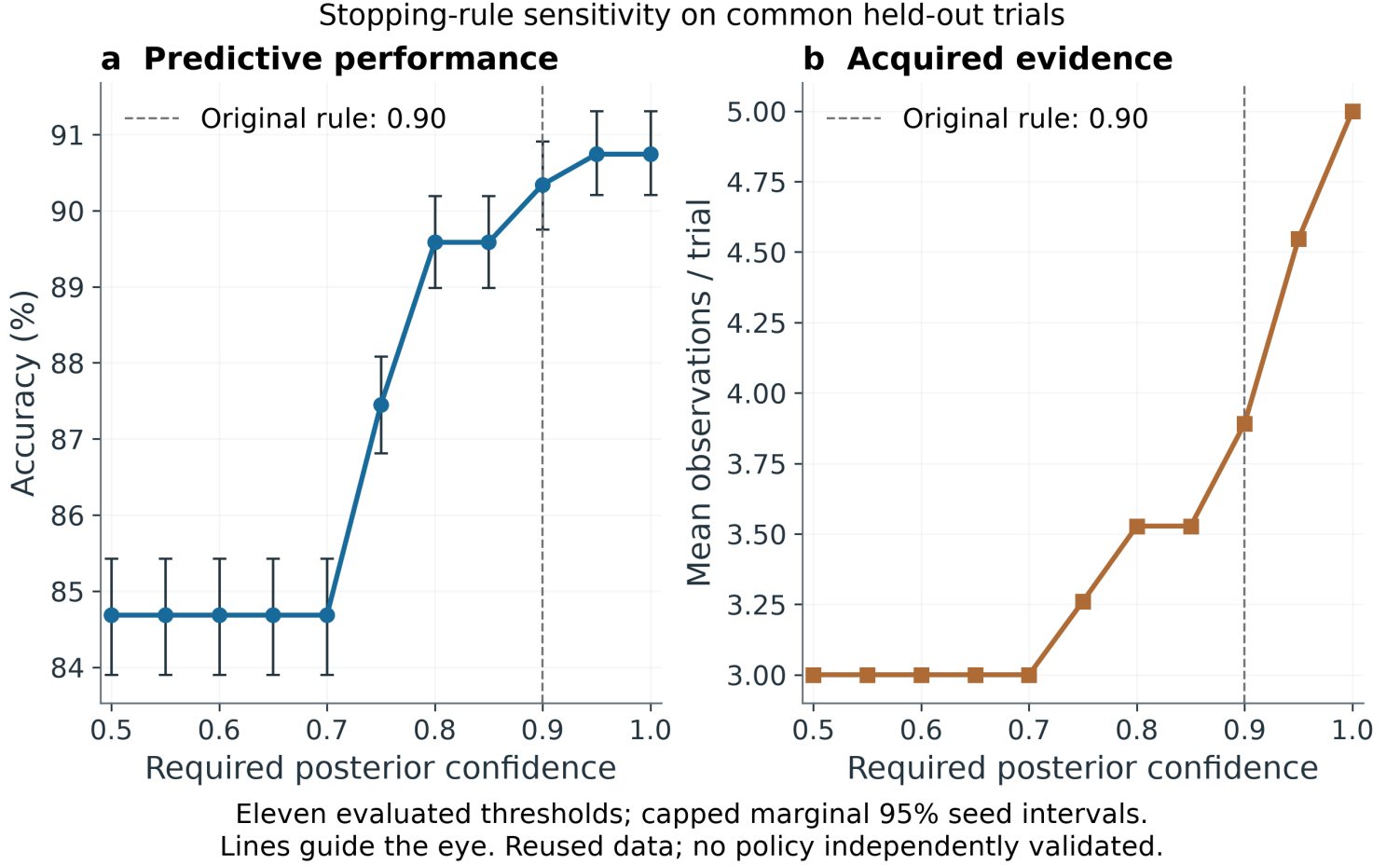


Figure 10: The stopping threshold jointly changes prediction and sensing resources. Eleven required posterior-confidence thresholds are evaluated using the same retained states and potential reports. (a) Accuracy and (b) observation count show seed means with capped marginal 95 percent bootstrap intervals; markers denote evaluated thresholds, and connecting lines only guide the eye. The dashed reference identifies the original 0.90 rule. Higher thresholds can acquire additional reports up to the five-observation budget. This reused-data sensitivity analysis neither independently validates a selected threshold nor assigns a price to predictive errors or observation cost.

There is no application-independent exchange rate between an accurate decision and an additional observation. A deployment loss could price each observation and each decision error, but a different price could change the preferred rule. The threshold sweep deliberately reports both outcomes rather than selecting a utility function after seeing the results. Label acquisition for Q4, knowledge of reliability for oracle conditions, and computational overhead remain additional resources outside the per-trial sensing count.

The exported calibration tables repeat the descriptive binning at 5, 10, and 20 bins, preserving counts and represented seeds in every occupied bin. Empty bins are omitted rather than assigned an invented observed frequency. Greater visual agreement at one resolution would not establish calibration at all resolutions, in small subgroups, or after environmental change. These diagnostics make uncertainty inspectable without turning a particular display convention into a scientific guarantee.

11.4.2 Executed EFE policy selection and its exact model expectations

The new action task distinguishes information quality from the preference for a null observation. Figure 11 shows the exact crossover of the two action scores and the finite-precision policy posterior. The state representation includes observation mode,

so the likelihood remains fixed while actions alter the predicted state distribution through B . This avoids attributing action-dependent sensing to a likelihood matrix whose meaning changes outside the declared model.

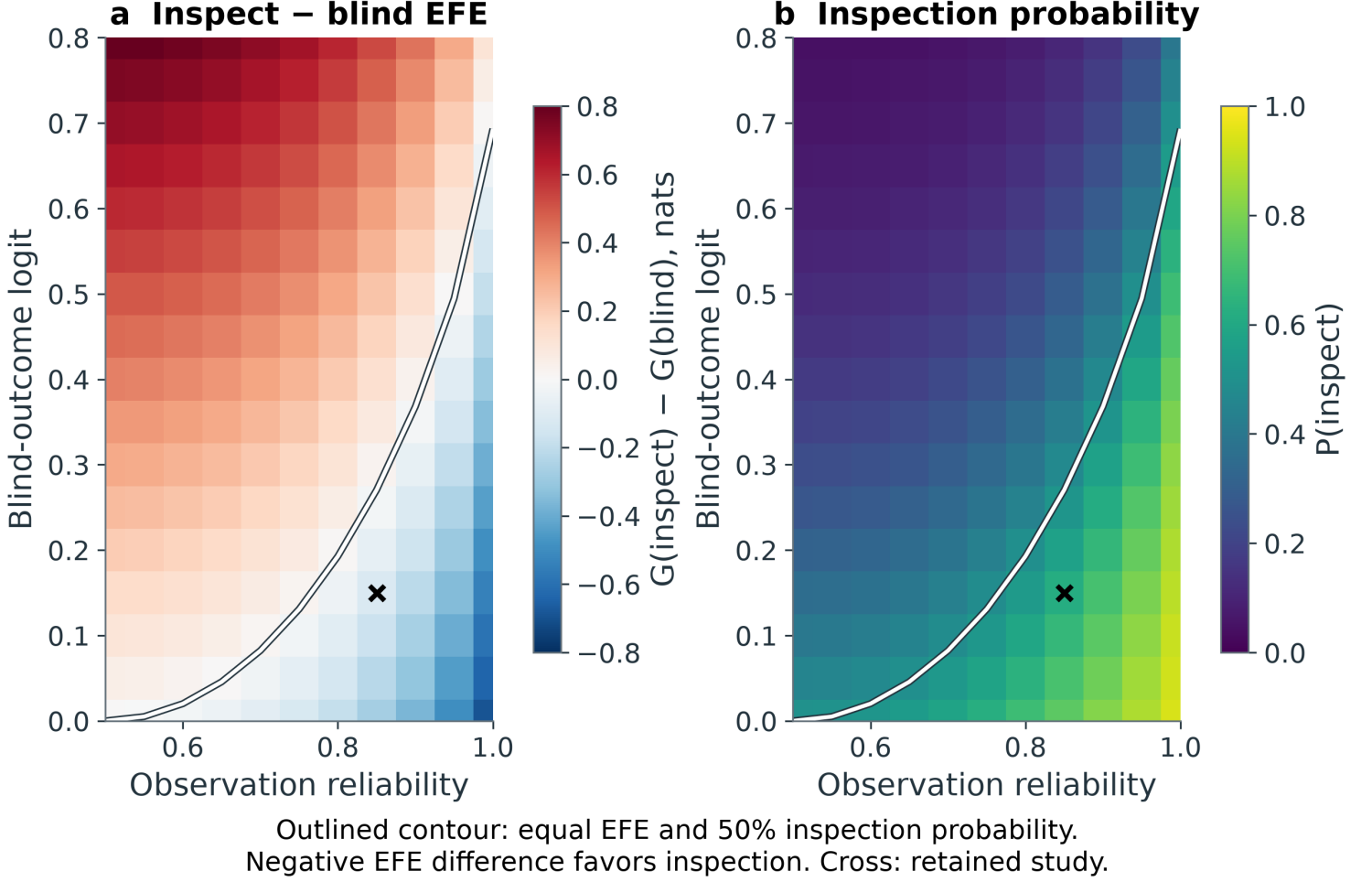


Figure 11: Exact EFE policy consequences of changing information quality and outcome preference. The 187 cells represent separately materialized GNN reliability/logit combinations. (a) Inspection minus blind normalized expected free energy is in nats; negative values favor inspection. (b) The corresponding inspection probability uses policy precision 4. The outlined white contour denotes equal EFE and hence probability one half for these two policies. The cross locates the retained sampled design. Actions change observation mode while preserving context. These are model evaluations, not independent experimental replicates; they have no sampling intervals.

Figure 5 compares the sampled policy with its fixed baselines and analytical expectations. A policy that sometimes chooses the preferred null observation can obtain fewer informative observations and lower context accuracy than always inspecting. EFE therefore need not maximize classification accuracy alone. The preference vector and policy precision help explain that behavior; they are substantive commitments, not estimated facts about a biological agent.

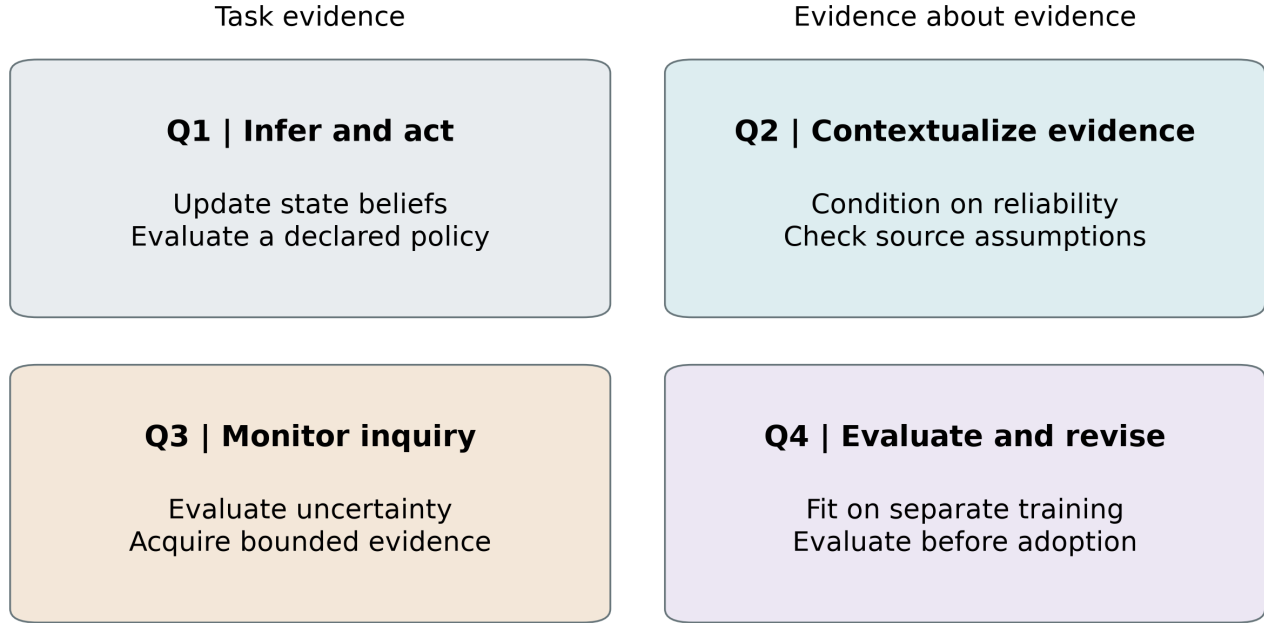
The sampled action outcomes are displayed in the main paper, Figure 5.

The exact reference expectations separate Monte Carlo variability from a change in the theoretical prediction. They are conditional on the same known generative model, so agreement supports computational consistency within this task rather than model adequacy in the world. Paired estimates and all seven metrics remain in the retained evidence, including resource and preference quantities that an accuracy-only ranking would omit.

In the retained action run, sampled EFE achieves 71.23% context accuracy with a marginal 95% interval of [70.23, 72.28]%, using 0.617 informative observations per trial with interval [0.607, 0.626]. The exact model predicts 71.64% accuracy and an inspection probability of 0.6182. Always inspecting achieves 84.90% accuracy, while always taking the blind action achieves 49.97%. These values describe the declared one-step task and its chosen preferences; they do not establish superiority of one policy for every objective.

11.5 Model revision mechanics and the limits of integrity diagnostics

a Four overlapping responsibilities



b From specification to justified revision

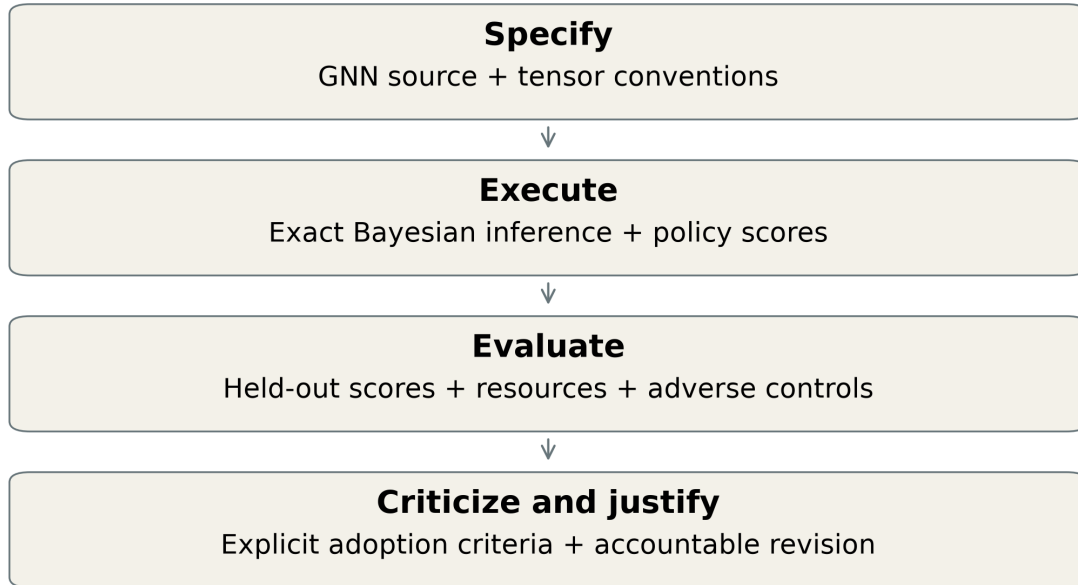


Figure 12: Inquiry roles and research procedures provide complementary descriptions of model work. (a) Q1 infers and acts; Q2 contextualizes evidence; Q3 monitors inquiry; Q4 evaluates and revises assumptions. Their responsibilities overlap. (b) Specification, execution, held-out evaluation and justified criticism form a proposed workflow. Arrows indicate procedural relations, not measured neural or causal connections. The experiments implement a bounded computational evaluation cycle and a supervised fitted-model comparison. Public participation, the authority to alter ends and open-ended autonomous revision remain outside the implemented mechanism.

An alternative weighted score $\tilde{G} = w_e(-I) + w_p(-\mathbb{E}[C]) + w_m K(m)$ can be useful as a designed decision objective. Arbitrary weights and an added metadata cost do not automatically preserve a variational derivation. Such scores must be labeled heuristics unless a corresponding generative model and derivation are supplied.

11.5.1 Diagnostics and their limits

`CognitiveSecurityAnalyzer.detect_anomaly` requires valid normalized belief distributions before applying a disclosed smoothing epsilon to KL divergence. The smoothing makes the diagnostic finite when a baseline assigns zero mass, and therefore it is not the unsmoothed support-sensitive KL used in the exact inference identities. Thresholds are user-specified and are not calibrated false-positive rates. Distribution shift can be benign, and an attack can remain close to a baseline.

`validate_framework_integrity` checks shape, finiteness, compatibility, nonnegativity, and stochasticity. A malicious model can satisfy every one of these contracts. Structural validity therefore cannot establish absence of tampering. The result explicitly records that tamper absence was not verified. The attack-surface scores and Gaussian parameter-drift demonstrations are author-specified illustrations, not measured risk scores or probability-preserving attack simulations.

11.6 Derivations, retained results, and reproduction

11.6.1 Algebraic derivation checks and analytical reference cases

Expanding $D_{\text{KL}}(q_o \| p_C)$ gives $-H(q_o) - \mathbb{E}[\log p_C]$. Adding ambiguity $\mathbb{E}_{q_s} H(A_{:,s})$ gives $-I(s; o) - \mathbb{E}[\log p_C]$, establishing Equation (3). Replacing normalized log preferences with logits subtracts $\log \sum_o \exp(C_o)$ per policy step. This constant affects reported magnitudes but not a fixed-model, fixed-convention policy ranking.

For VFE, substitute $p(o, s) = p(s | o)p(o)$ into Equation (2). The log-evidence term factors out of the normalized expectation, leaving posterior KL divergence. With $q = p(s | o)$ the divergence is zero. Exact-zero handling uses $0 \log 0 = 0$ and retains infinite divergences for support violations.

Tests additionally compare batched binary log-odds inference with sequential core Bayesian updates, check asymmetric column-stochastic transitions, and check preference-shift and horizon-aggregation identities. Mathematical identity tests complement software coverage; coverage alone does not establish theoretical correctness.

11.6.2 Retained sensor-study estimates and primary paired contrasts

Condition	Accuracy (%)	Brier loss	Log loss (nats)	Observations/trial
Q1_fixed	76.36	0.1781	0.5519	3.000
Q2_metadata	84.68	0.1235	0.4051	3.000
Q3_adaptive	90.34	0.0812	0.2896	3.892
Q4_learned	84.68	0.1258	0.4114	3.000
Q2_misleading	54.12	0.3306	1.1861	3.000
Fixed_four	86.75	0.0924	0.3017	4.000
Fixed_five	90.74	0.0708	0.2361	5.000

Contrast	Difference (points)	95% interval	d_z	Holm p
H1	8.325	[7.442, 9.217]	3.439	0.00030
H2	5.658	[5.058, 6.267]	3.275	0.00030
H3	8.325	[7.442, 9.217]	3.439	0.00030

These estimates are computed from `results/summary.json`. Accuracy differences are in percentage points. Each comparison uses 30 paired seed means. The primary p-values are Monte Carlo sign-flip estimates at finite resolution, with Holm correction across H1–H3. The intervals are marginal percentile-bootstrap intervals; they are not adjusted simultaneous intervals. The supplementary paired t-tests and full seed-level data are retained in the artifact set.

11.6.3 Active-sensing outcomes and the distinction between scores and preferences

Action strategy	Accuracy (%)	Brier loss	Log loss (nats)	Inspections/trial
Sampled EFE	71.23	0.1747	0.5270	0.617
Always inspect	84.90	0.1282	0.4244	1.000
Always blind	49.97	0.2500	0.6931	0.000

These are means over 30 paired seeds, each with 200 generated world trials. The action world is sampled from the actual imported GNN D/B/A tensors and every posterior uses the same declared numerical contracts. The 187-point exact policy grid

has a maximum risk-plus-ambiguity identity residual of 1.11×10^{-16} . Full marginal intervals, exact expectations and 14 paired comparisons are retained in the active-sensing CSV tables. These exploratory analyses introduce no additional hypothesis-test family.

The minimum-EFE action and the sampled action have different meanings. At the declared precision, the policy posterior retains stochastic choice. Always inspecting is a comparator with a larger informative-sensing budget, rather than an algorithm that estimates or optimizes the supplied preference vector. Thus, the results do not establish universal superiority, autonomous goal revision or empirical validity of the environment model.

11.6.4 Modular implementation and asymptotic computational complexity

core defines categorical models and exact inference. **framework** supplies the taxonomy, monitoring illustrations, and structural security diagnostics. **analysis** contains generators, statistics, validation, experiments, and pipeline orchestration. **visualization** creates educational and experiment figures; **utils** supplies manuscript validation and provenance. No parent repository infrastructure is required by the research pipeline.

For a dense model with N_s states, N_o outcomes, N_a available actions and a policy of T steps, a single policy costs approximately $\mathcal{O}(T(N_s^2 + N_o N_s))$ for transitions and predictions, with $\mathcal{O}(N_o N_s)$ likelihood entropy preprocessing. Evaluating N_π candidate policies multiplies the per-policy cost by N_π . The available-action count identifies the transition tensor’s third axis; it is not an additional factor when evaluating a fixed sequence. The implementation does not claim sparse scaling or measured support for 1,000-state models. No runtime table is reported without a retained benchmark protocol and hardware description.

11.6.5 Reproduction with pinned GNN and publication-renderer dependencies

Run from a clean checkout:

First provide the real upstream GNN checkout through `GNN_ROOT`, pinned by `src/core/gnn_models/upstream.lock.json`. The supported package minimum is Python 3.11; the recorded scientific environment is Python 3.12. No upstream pipeline or native backend is silently substituted for the local exact categorical runtime.

```
uv sync --locked --extra dev --extra research --python 3.12
MPLBACKEND=Agg uv run --python 3.12 pytest --cov=src --cov-report=term-missing
uv run --python 3.12 python scripts/07_analysis_pipeline.py
```

The pipeline writes the source-backed results and standalone HTML manuscript to `results/`. Changing seed counts, trial counts, training size, numerical libraries or model configuration creates a different run, which must be described as such. Random-number generators are local; reproducibility does not rely on global `numpy.random.seed`. Source and output digests support verification within a recorded environment. PDF figure metadata or library/platform differences can change binary hashes without changing numerical estimates.

The older scripts 01–05 generate educational diagrams under `../figures/`; those illustrations are separate from the retained experiment evidence. Script 06 validates the manuscript rather than inserting unreviewed or stale figures automatically.

12 Supplemental Notation for Models, Objectives, and Statistics

This is the common notation reference for the main paper and technical supplement. Symbols denote quantities within a specified model; they do not by themselves identify a biological mechanism or justify a normative objective. The tensor atlas in Figure 7 shows the concrete conventions used by the packaged models.

12.1 States, observations, and executable model components

Symbol	Meaning and convention
s, s'	Current and next hidden-state indices; the first transition axis is the destination
o	Categorical observation or outcome index
a	Action index; action 0 is blind and action 1 is inspect in the action study
j	Dummy summation index over the state or outcome set indicated by the expression
m	Metadata about an observation's source, reliability, context or provenance
$A, A_{o,s}$	Observation likelihood $p(o s)$; columns sum to one
$A_{:,s}$	Full outcome distribution conditional on state s ; colon selects all outcomes
$B, B_{s',s,a}$	Transition likelihood $p(s' s, a)$; each action slice is column stochastic
$B(a)$	Matrix slice of B for action a ; multiplying a state column vector predicts its successor
C, C_o	Finite outcome preference logits, not probabilities or measured rewards
D, D_s	Normalized initial state prior
$p_C(o)$	Softmax-normalized outcome preference induced by C
$p(o, s), p(o), p(s o)$	Joint model probability, evidence and exact conditional posterior
$p(o s, m)$	Observation likelihood conditioned on metadata
$q, q(s), q_s$	Candidate, posterior or predictive state distribution, specified by context
q_o	Predictive outcome distribution
$e(o)$	Normalized nonnegative virtual-evidence weights over possible observations
$L_e(s)$	Evidence likelihood induced by the weighted sum of observation likelihoods
r	Actual informative-observation reliability; r_i is the reliability of sensor i
ρ	Agent-assumed informative-observation reliability in the mismatch analysis
c	Blind/null-outcome preference logit in the action task
w	World subscript; distinguishes world probabilities/tensors from agent assumptions
$p_w(s', o a)$	World joint distribution of next state and outcome given action
$y(s')$	Binary context label given by next-state parity in the four-state action model
$\hat{p}_a(o)$	Agent forecast of context 1 after outcome o under action a

The action model orders states as blind context 0, blind context 1, informative context 0, informative context 1. Outcomes are null, indication of context 0, indication of context 1. The diagram's labels s0–s3 denote those state indices, not four state factors. The single-factor GNN profile admits a categorical state variable; the product interpretation is an authored meaning for its categories.

12.2 Information measures, evaluative objectives, and policy selection

Symbol	Meaning and convention
$F(q, o)$	Exact variational free energy for a candidate posterior and explicit joint model
G, G_τ	Normalized expected free energy, or its contribution at policy step τ
$G_{\text{inspect}}, G_{\text{blind}}$	Normalized EFE for the two one-step sensing policies
π	An open-loop action sequence when used as a policy; $\pi(a)$ denotes a one-step action distribution in the world-risk expression
τ, q_τ, a_τ	Policy step, predicted state distribution and prescribed action at that step
γ	Positive policy precision multiplying negative EFE before normalization
$H(\cdot)$	Shannon entropy; distinct from the horizon notation used in some external literature
$h(r)$	Binary entropy at reliability r
$I, I(s; o \pi)$	Expected state information gain under the specified policy
$D_{\text{KL}}[q p]$	KL divergence; infinite when q gives positive mass where p is zero
$\mathbb{E}_q[\cdot]$	Expectation under distribution q ; omitted subscripts are specified locally
$\log, \exp$	Natural logarithm and exponential; information quantities use nats
$\ell(\hat{p}, y)$	Binary scoring loss applied to a forecast and realized context label
$R_\ell(\pi; w, \text{agent})$	Exact expected scoring loss under world probabilities and an agent's forecasts
$R_{\text{Brier}}, R_{\log}$	Exact Brier and logarithmic risks
$r(o), \beta$	Reward function and positive scale in the related-work mapping from reward to preference; $r(o)$ is not sensor reliability
$\tilde{G}$	Illustrative weighted decision score, not the executed normalized EFE
w_e, w_p, w_m	Authored heuristic weights for information, preference and metadata terms
$K(m)$	Authored metadata penalty in that illustrative weighted score

total is the compatibility API's raw-logit score; **normalized_total** is its normalized counterpart. Their fixed-model relationship and horizon conventions are stated in Supplement 11.2. Adding a constant to all preference logits leaves normalized preferences unchanged. These API fields are not interchangeable when comparing models with different preference normalizers.

12.3 Statistical estimands, uncertainty, and computational notation

Symbol or term	Meaning and convention
$\Delta, \overline{\Delta}, s_{\Delta}$	Paired seed-level difference, its mean and its sample standard deviation
$d_z = \overline{\Delta}/s_{\Delta}$	Standardized within-pair difference; undefined for zero variation unless explicitly handled
H1–H3	The three primary paired accuracy comparisons; their exact contrasts are in the protocol
p -value	Tail evidence from the named test; distinct from model probabilities written as $p(\cdot)$
Beta(1,1)	Uniform prior over a sensor reliability in the supervised fitting procedure
N_s, N_o, N_a	Counts of hidden states, outcomes and available actions
N_{π}, T	Number of evaluated candidate policies and number of steps in each policy
$\mathcal{O}(\cdot)$	Asymptotic dense arithmetic cost, not an observed runtime benchmark
Seed	Independent generated-dataset unit for paired uncertainty estimation
Trial	A held-out context and potential observations within a seed
Marginal interval	An interval for one stated quantity, without simultaneous familywise coverage
Sign flip	Paired resampling test requiring the stated sign-exchangeability assumption

12.4 Conceptual terms and epistemic boundaries

Q1–Q4 name inferential, contextual, monitoring and revision roles in Figure 1; they are not measured natural cognitive kinds. **Metadata** is information about the source, production or reliability of task-level evidence. Its role depends on the task relation.

Monitoring evaluates a property of inference; **control** changes processing on that basis. **Calibration** concerns agreement between probabilistic forecasts and observed frequencies, rather than posterior concentration. Proper losses measure forecast quality and need not be pure calibration statistics.

Meta-epistemic practice examines the representation, acquisition and reliability of evidence. **Meta-pragmatic practice** examines action, preferred consequences, budgets and the criteria authorizing revision. **Structural integrity** denotes validity of numerical contracts; it does not establish truthful metadata, absence of tampering or legitimacy of preferences.

13 References

Publication records and preprints are identified separately below. The scholarship register records source identities, reading scope, and which interpretations each added reference supports.

Active Inference Institute. Generalized notation notation: Executable generative model specification, 2026. URL https://github.com/ActiveInferenceInstitute/Generalized_Notation_Notation/blob/46459949944e3217f8a6a804d4ff8c52829ef275/docs/gnn/reference/gnn_syntax.md. Software and normative syntax, pinned revision 46459949944e3217f8a6a804d4ff8c52829ef275; reference parser imported by this study.

Jane Addams. *Democracy and Social Ethics*. The Macmillan Company, New York, 1902. URL <https://www.gutenberg.org/files/15487/15487-h/15487-h.htm>. Primary text; especially Introduction and Charitable Effort.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL <https://arxiv.org/abs/1606.06565>.

Majid D. Beni and Ahti-Veikko Pietarinen. Aligning the free-energy principle with Peirce’s logic of science and economy of research. *European Journal for Philosophy of Science*, 11:94, 2021. doi: 10.1007/s13194-021-00408-y. URL <https://link.springer.com/article/10.1007/s13194-021-00408-y>. Article 94.

M. Bergamaschi Ganapini, M. Campbell, F. Fabiano, L. Horesh, J. Lenchner, A. Loreggia, N. Mattei, F. Rossi, B. Srivastava, and K. B. Venable. Fast, slow, and metacognitive thinking in AI. *npj Artificial Intelligence*, 1:27, 2025. doi: 10.1038/s44387-025-00027-5. URL <https://www.nature.com/articles/s44387-025-00027-5>.

- Jochen Bröcker and Leonard A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007. doi: 10.1175/WAF993.1. URL https://www.personal.reading.ac.uk/~pt904209/publications/reliable_reldiags.pdf. Author manuscript dated November 2006; journal publication 2007.
- COPE Council. Authorship and AI tools. Cope position, Committee on Publication Ethics, 2024. URL <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools>. Version 1 published August 20, 2024; position last reviewed February 13, 2023.
- 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. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC7732703/>.
- Scott David, Richard J. Cordes, and Daniel A. Friedman. Active inference in modeling conflict, 2021. URL <https://zenodo.org/records/5759807>. Preprint, version v1.2, December 2; especially pp. 17–18 and 37–38, Figure 9.
- A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856>.
- Bert de Vries, Wouter Nuijten, Thijs van de Laar, Wouter Kouw, Sepideh Adamiat, Tim Nisslbeck, Mykola Lukashchuk, Hoang Minh Huu Nguyen, Marco Hidalgo Araya, Raphael Tresor, Thijs Jenneskens, Ivana Nikoloska, Raaja Ganapathy Subramanian, Bart van Erp, Dmitry Bagaev, and Albert Podusenko. Expected free energy-based planning as variational inference, 2025. URL <https://arxiv.org/abs/2504.14898>. Preprint, version 4.
- Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL <https://jmlr.org/papers/v7/demsar06a.html>.
- John Dewey. The reflex arc concept in psychology. *Psychological Review*, 3(4):357–370, 1896. doi: 10.1037/h0070405. URL <https://psychclassics.yorku.ca/Dewey/reflex.htm>.
- John Dewey. *How We Think*. D. C. Heath and Co., Boston, New York and Chicago, 1910. URL <https://www.gutenberg.org/files/37423/37423-h/37423-h.htm>. Original 1910 text, Chapter I; distinct from the 1933 revision.
- John Dewey. *Reconstruction in Philosophy*. Henry Holt and Company, New York, 1920. URL <https://www.gutenberg.org/files/40089/40089-h/40089-h.htm>. Primary text; especially Chapters VI and VII.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993. doi: 10.1007/978-1-4899-4541-9. URL <https://www.hms.harvard.edu/bss/neuro/bornlab/nb204/statistics/bootstrap.pdf>.
- John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10):906–911, 1979. doi: 10.1037/0003-066X.34.10.906. URL <https://doi.org/10.1037/0003-066X.34.10.906>.
- Stephen M. Fleming and Hakwan Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014. doi: 10.3389/fnhum.2014.00443. URL <https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00443/full>.
- Daniel Friedman. Synthesis of agent and niche, 10 2025. URL <https://zenodo.org/records/17243488>. Version v2, October 1; speculative conceptual and literary work, Zenodo.
- Daniel Ari Friedman. Before pragmatism had a name: Blake’s America A Prophecy anticipates american anticipatory epistemology, 3 2026a. URL <https://zenodo.org/records/18984939>. Version v2, March 12; Zenodo publication record; conceptual scholarship.
- Daniel Ari Friedman. Prior cognitive art, 2026b. URL <https://zenodo.org/records/21316511>. Version 0.1.0; conceptual working paper, Zenodo; deterministic illustrations.
- Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010. doi: 10.1038/nrn2787. URL <https://www.nature.com/articles/nrn2787>.
- Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas FitzGerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015. doi: 10.1080/17588928.2015.1020053. URL <https://pubmed.ncbi.nlm.nih.gov/25689102/>.
- Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O’Doherty, and Giovanni Pezzulo. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, 68:862–879, 2016. doi: 10.1016/j.neubiorev.2016.06.022. URL <https://discovery.ucl.ac.uk/id/eprint/1506130/>.
- Karl Friston, Conor Heins, Tim Verbelen, Lancelot Da Costa, Tommaso Salvatori, Dimitrije Markovic, Alexander Tschantz, Magnus Koudahl, Christopher Buckley, and Thomas Parr. From pixels to planning: scale-free active inference. *Frontiers in Network Physiology*, 5:1521963, 2025. doi: 10.3389/fnetp.2025.1521963. URL <https://www.frontiersin.org/journals/network-physiology/articles/10.3389/fnetp.2025.1521963/full>.
- Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020. doi: 10.1007/s11023-020-09539-2. URL <https://link.springer.com/article/10.1007/s11023-020-09539-2>.

- Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, 2013. doi: 10.1111/j.2044-8317.2011.02037.x. URL <https://stat.columbia.edu/~gelman/research/published/philosophy.pdf>.
- Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow, 2020. URL <https://arxiv.org/abs/2011.01808v1>. Preprint, version 1; distinct from the later book.
- Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011. doi: 10.1146/annurev-psych-120709-145346. URL <https://www.annualreviews.org/content/journals/10.1146/annurev-psych-120709-145346>. A separate erratum is recorded at DOI 10.1146/annurev-ps-62-040111-200001.
- Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL <https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf>.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. URL <https://proceedings.mlr.press/v70/guo17a.html>.
- Conor Heins, Beren Millidge, Lancelot Da Costa, Stephen Mann, Michael Kirchhoff, Magnus Koudahl, Alexander Tschantz, Christopher L Buckley, and Karl Friston. 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. URL <https://joss.theoj.org/papers/10.21105/joss.04098>.
- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL <https://www.jstor.org/stable/4615733>.
- William James. *Pragmatism: A New Name for Some Old Ways of Thinking*. Longmans, Green, and Co., 1907. URL <https://www.gutenberg.org/ebooks/5116>. Primary text; especially Lectures II and VI. Electronic transcription at Project Gutenberg.
- Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020. doi: 10.1017/S0140525X1900061X. URL https://cocosci.princeton.edu/papers/lieder_resource.pdf.
- George Herbert Mead. *Mind, Self, and Society from the Standpoint of a Social Behaviorist*. University of Chicago Press, Chicago, 1934. URL https://brocku.ca/MeadProject/Mead/pubs2/mindself/Mead_1934_toc.html. Posthumous edited volume; primary-text Sections 16 and 24 consulted.
- Beren Millidge, Alexander Tschantz, and Christopher L Buckley. Whence the expected free energy? *Neural Computation*, 33(2): 447–482, 2021. doi: 10.1162/neco_a_01354. URL <https://arxiv.org/abs/2004.08128>.
- Thomas O Nelson and Louis Narens. Metamemory: A theoretical framework and new findings. In *Psychology of Learning and Motivation*, volume 26, pages 125–173. Academic Press, 1990. doi: 10.1016/S0079-7421(08)60053-5. URL https://www.imbs.uci.edu/~lnarens/1990/Nelson%26Narens_Book_Chapter_1990.pdf.
- J. Pazem, M. Krumm, A. Q. Vining, L. J. Fiderer, and H. J. Briegel. Free energy projective simulation (FEPS): Active inference with interpretability. *PLOS One*, 20(9):e0331047, 2025. doi: 10.1371/journal.pone.0331047. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0331047>.
- Charles S. Peirce. The fixation of belief. *Popular Science Monthly*, 12:1–15, 1877. URL <https://www.peirce.org/writings/p107.html>. November; primary-text transcription.
- Charles S. Peirce. How to make our ideas clear. *Popular Science Monthly*, 12:286–302, 1878. URL <https://www.peirce.org/writings/p119.html>. January; primary-text transcription.
- Charles S. Peirce. What pragmatism is. *The Monist*, 15(2):161–181, 1905. URL https://cspeirce.com/menu/library/bycsp/what_is/whatpragis.htm. April; primary-text transcription with Collected Papers paragraph numbers.
- 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://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833>.
- Noor Sajid, Philip J Ball, Thomas Parr, and Karl J Friston. Active inference: demystified and compared. *Neural Computation*, 33(3):674–712, 2021. URL <https://arxiv.org/abs/1909.10863>.
- Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955. doi: 10.2307/1884852. URL <https://doi.org/10.2307/1884852>.

- Ryan Smith, Karl J Friston, and Christopher J Whyte. A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*, 107:102632, 2022. doi: 10.1016/j.jmp.2021.102632. URL <https://pmc.ncbi.nlm.nih.gov/articles/PMC8956124/>.
- Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637. URL <https://doi.org/10.1177/1745691616658637>.
- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration, 2018. URL <https://arxiv.org/abs/1804.06788>. Preprint.
- N. Tinbergen. On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, 20(4):410–433, 1963. doi: 10.1111/j.1439-0310.1963.tb01161.x. URL <https://onlinelibrary.wiley.com/doi/10.1111/j.1439-0310.1963.tb01161.x>.
- Alexander Tschantz, Manuel Baltieri, Anil K. Seth, and Christopher L. Buckley. Scaling active inference, 2019. URL <https://arxiv.org/abs/1911.10601>. Preprint.
- Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467, 2019. URL <https://proceedings.mlr.press/v89/vaicenavicius19a.html>.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. doi: 10.1007/s11222-016-9696-4. URL <https://arxiv.org/abs/1507.04544>.
- Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. MetaMind: Modeling human social thoughts with metacognitive multi-agent systems. *arXiv preprint arXiv:2505.18943*, 2025. URL <https://arxiv.org/abs/2505.18943>. Preprint; cited as architectural context, not active-inference validation.