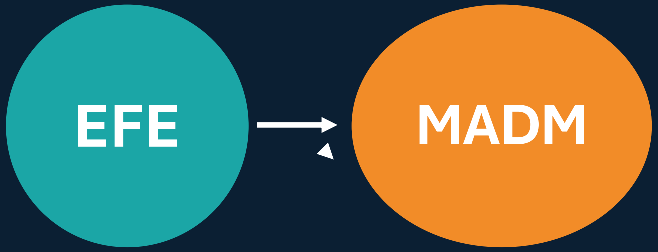
Active Inference as Multi-Attribute Decision Making

Explainable Planning via Expected Free Energy Decomposition

28th International Conference on Artificial Intelligence



Decisions can be explained by the same decomposed quantities that select actions. : Expected Free Energy can be interpreted as a dynamic multi-attribute decision objective, making Active Inference explainable by construction.

Bhagyeshkumar Chokhawala and Dr. Atif Farid Mohammad
Capitol Technology University

The Explainability Gap in Sequential AI

High-performing policies often lack faithful, inspectable decision structure.

Deep RL / Neural Policy optimizes scalar reward → **Post-hoc XAI approximates explanation** → **Need explanation by construction**

?

Scalar reward collapses risk, uncertainty, preferences, and safety into one number.

≈

Local explanations may not match the objective that actually selected the action.

✓

Decision attributes should remain visible before, during, and after action selection.

Core gap

Most systems optimize a scalar reward. That collapses multiple considerations into one number and makes trade-offs hard to inspect.

In long-horizon settings, good explanations should expose:

- uncertainty and hidden-state beliefs
- preference or goal alignment
- predicted outcomes and trade-offs
- why one action was preferred over alternatives

Key requirement: preserve multi-attribute trade-offs through the planning objective itself.

Active Inference and Expected Free Energy

A unified objective for perception, learning, and action under uncertainty.

Observation o_t → **Belief update** $q(s_t)$ → **Generative prediction** $p(o,s)$ → **Expected Free Energy** $G(\pi)$ → **Action / policy** π^* → **Outcome** $o_{(t+1)}$

ACTIVE INFERENCE STACK

- **Observations** — sensory evidence o_t
- **Belief Updating** — infer latent state $q(s)$
- **Generative Model** — predict outcomes $p(o,s)$
- **Policy Evaluation** — compute EFE $G(\pi)$
- **Action** — execute policy $\pi(a|s)$

MODEL COMPONENTS

- **Prior Preferences C** — preferred outcomes
- **Likelihood A** — states generate observations
- **Transition B** — actions change states

DECISION LOOP

1. receive observation
2. update beliefs over latent states
3. simulate future outcomes
4. evaluate EFE
5. select action minimizing surprise
6. act and observe again

EFE COMPONENTS

Epistemic info gain	**Pragmatic** goal alignment
Ambiguity uncertainty	**Risk** divergence

EFE(policy) = weighted trade-off over these components

Beliefs encode partial observability; actions minimize predicted uncertainty and preference divergence.

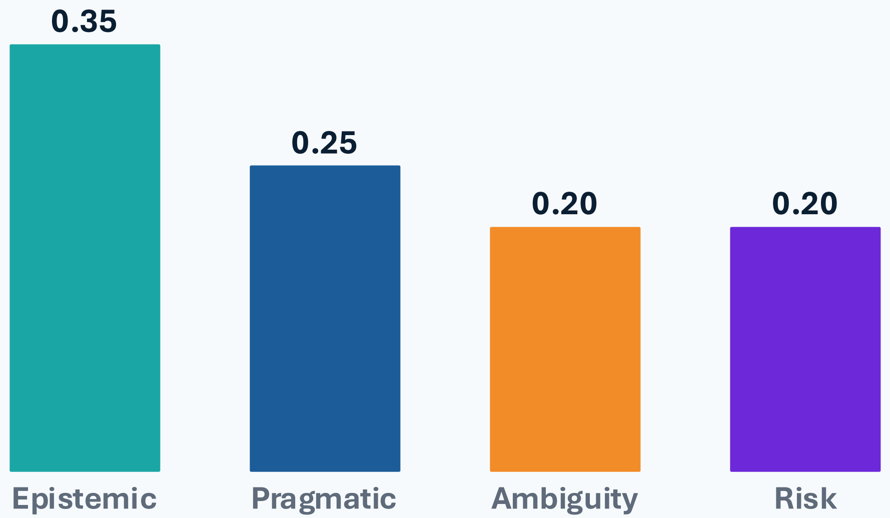
Exploration = epistemic value

Exploitation = pragmatic value

Decision = minimize EFE

Expected Free Energy Decomposes into Decision Attributes

The decision objective is already multi-attribute.



Active Inference unifies exploration and exploitation under one variational objective.

EFE encourages actions that:

- Reduce uncertainty about hidden states
- Align outcomes with prior preferences
- Avoid ambiguous or risky futures

This makes sequential planning uncertainty-aware and preference-aware at the same time.

Information Gain: Epistemic
information gain
ask: what will I learn?

Goal / preference Alignment: Pragmatic
preference alignment
ask: what do I want?

Outcome uncertainty: Ambiguity
uncertainty in outcomes
ask: how unclear is the future?

Safety / risk Sensitivity: Risk
divergence from preferred outcomes
ask: what can go wrong?

Instead of adding interpretability later, the EFE objective exposes the criteria driving each action.

From EFE Decomposition to Adaptive MADM

EFE minimization has the same structure as weighted multi-criteria optimization.

Classical MADM

$\text{Score}(a) = \Sigma\, w_k \cdot c_k(a)$

Alternatives are scored across multiple criteria using weights. Typically, criteria are handcrafted and weights are static.

Active Inference

$G(a) = \Sigma\, w_k(s) \cdot g_k(a)$

EFE already decomposes into multiple criteria. Beliefs make the effective weights dynamic and state-dependent.

c_1 / g_1	uncertainty reduction
c_2 / g_2	preference alignment
c_3 / g_3	ambiguity minimization
c_4 / g_4	risk sensitivity

Decision: $a^* = \arg\min G(a)$

What changes versus classical MADM?

Weights are not fixed. Belief updates and experience produce state-dependent attribute salience.

Structural equivalence:

Minimizing EFE is equivalent to minimizing a weighted sum of attribute-level contributions.

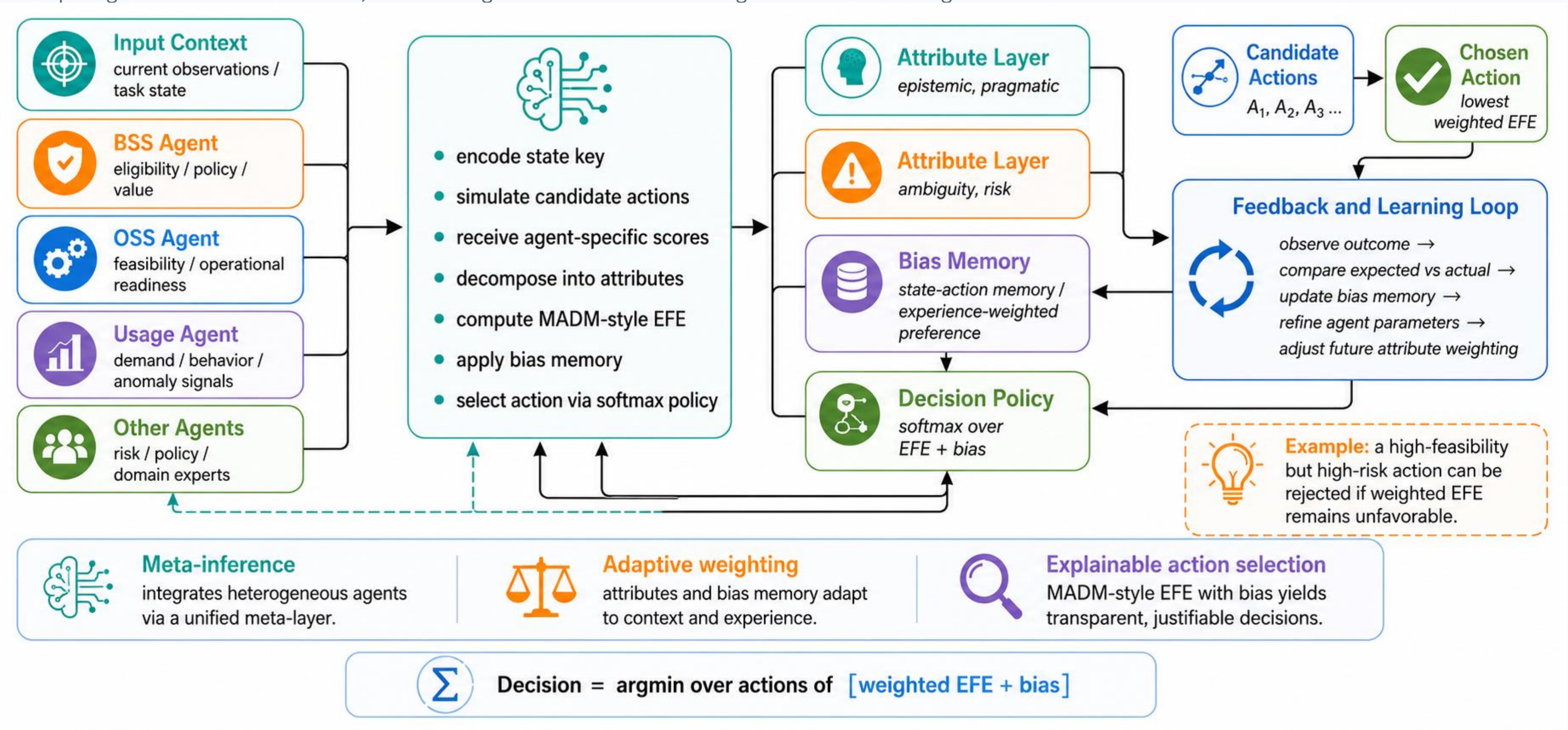
Key difference from classical MADM:

- weights evolve with belief updates and context
- decision criteria remain explicit and inspectable

Core claim: Active Inference operationalizes MADM through decomposable EFE and dynamic weighting.

PlanningEFEMix: MADM-Aware Meta-Planning Architecture

Multiple agents contribute evidence; the meta-agent selects actions through MADM-EFE scoring.



Illustrative MADM Decision Example

Illustrative activation decision under partial observability.

Attribute	Source	Weight	EFE	Contribution
Business validity	Eligibility / pricing	0.30	0.15	0.045
Operational feasibility	Network Readiness	0.45	0.62	0.279
Demand / usage signal	Demand / anomaly signal	0.25	0.38	0.095
Total MADM-EFE				**0.419**

Candidate actions

Immediate activation	EFE 0.419	p 0.18
Deferred activation	**EFE 0.270**	**p 0.54**
Manual review	EFE 0.330	p 0.28

Selected Decision: Deferred activation

Explanation: feasibility uncertainty dominated short-term optimality in this state.

Formal Equivalence: Active Inference as Adaptive MADM

Theorem-level claim: minimizing EFE solves a dynamic weighted multi-attribute problem.

Theorem 1
Structural Equivalence

If Expected Free Energy decomposes into additive attribute terms, then minimizing EFE is equivalent to solving a MADM problem with state-dependent weights.

$G(a) = \Sigma\, g_k(a)$ → **Beliefs update** $w_k = f(q(s), \text{context})$ → $a^* = \arg\min \Sigma\, w_k(s)\, g_k(a)$

GUARANTEED PROPERTIES

Decomposability
EFE separates into additive attribute terms.

Causal faithfulness
Each term maps to a distinct decision attribute.

Contextual adaptivity
Weights shift with beliefs and context.

Traceability
The chosen action exposes every attribute contribution.

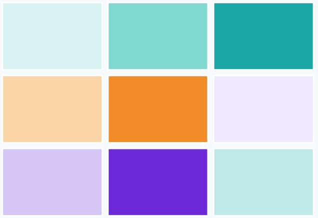
Intrinsic Interpretability Artifacts

Explanations are generated from the decision objective, not a surrogate model.

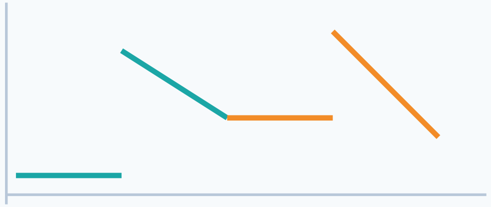
Attribute-Level EFE Table

Action	Risk	Amb	Pref
A1	0.4	0.2	0.1
A2	0.2	0.1	0.3

Bias Heatmap



Decision Trajectory Graph



Attribute-level EFE tables

Compare candidate actions and identify dominant attributes.

Useful for: identifying what dominated the final choice.

Bias heatmaps

Visualize learned state-action tendencies.

Useful for: understanding stable preferences and context specialization.

Decision trajectories

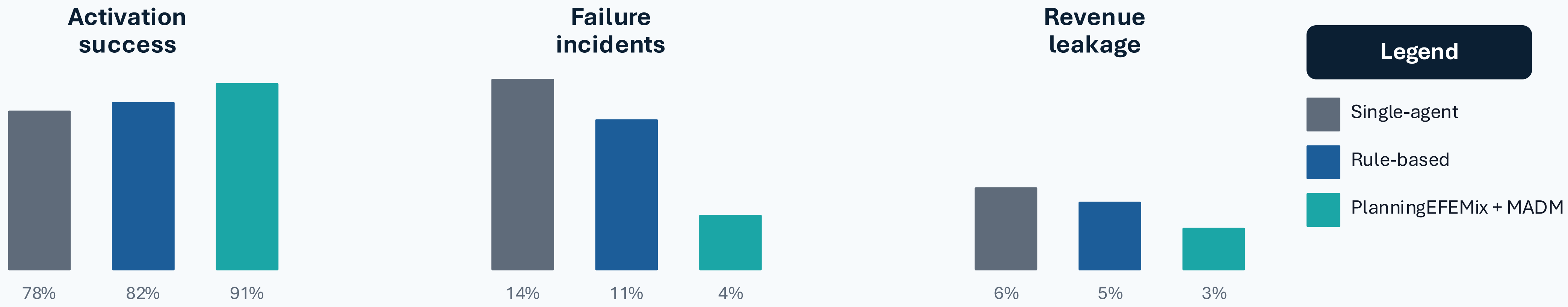
Track how attribute contributions evolve over time.

Useful for: showing shifts between exploration, exploitation, ambiguity reduction, and risk avoidance.

Auditability is supported at local decision, state-memory, and temporal-policy levels.

Experimental Results: MADM Improves Robust Decision Selection

Experimental simulation summary across 1,000 decisions.



vs. post-hoc explainability & scalar-reward RL, Active Inference provides:

- high explanation faithfulness
- high causal traceability
- dynamic trade-off management

Example evaluation dimensions:

- performance
- trade-off quality
- adaptivity
- counterfactual clarity

Comparison snapshot

Post-hoc attribution

- approximate feature effects
- lower faithfulness

Reward-decomposed RL

- partially interpretable
- still relies on manual scalarization

Active Inference

- EFE decomposition is intrinsic
- high faithfulness and traceability

Ablation insight: removing bias learning, softmax selection, or the agent grid reduces adaptation in ambiguous states.

Practical Implications and Industry Suitability

The same framework can adapt attribute priorities across domains.

HE

Healthcare

treatment planning, triage, safety-aware decisions

FI

Finance

fraud detection, portfolio rebalancing, risk control

MA

Manufacturing

fault tolerance, fault handling, process optimization

LO

Logistics

route feasibility, scheduling, uncertainty-aware fulfillment

EN

Energy

demand – supply balancing, resilience management

DI

Digital platforms

Personalization, personalization, adaptive control

General principle: feasibility, risk, uncertainty, preferences, and policy constraints can trade off dynamically as conditions change.

Limitations and Deployment Considerations

Translating the theory into practice requires architectural and governance safeguards.

1 Computational overhead

Multiple internal agents require memory and compute; use pruning, abstraction, distillation.

2 Feedback quality

Sparse or biased feedback can slow adaptation or skew weights.

3 Expert interpretation

Multi-attribute EFE trade-offs may require domain expertise.

4 Governance & safety

Human oversight, monitoring, and calibration of exploration vs. exploitation are critical in high-risk settings.

Scalability levers

- state abstraction
- hierarchical planning
- agent pruning
- model distillation
- cached inference
- parallel pipelines

Deployment requires balancing responsiveness, safety, interpretability, and learning stability.

Future Work

Scaling intrinsic explainability to high-dimensional, continuous, and real-world settings.

1
Continuous spaces

Deep latent-state inference and amortized EFE decomposition.

2
Visualization tooling

Standardized dashboards for high-dimensional attributes.

3
Counterfactual traces

Explain why selected actions beat rejected alternatives.

4
Hybrid ensembles

Extend PlanningEFEMix while preserving EFE explanations.

The research agenda is not only more accuracy, but better auditable decision-making.

Thank You

Research Profile QR Code

Scan the QR code to open my research profile, publications, Google Scholar, ORCID, and professional links.

Key Contributions

Active Inference as a foundation for inherently transparent sequential decision-making.

1. **Reframed Active Inference as adaptive MADM**
2. Formalized EFE minimization as weighted multi-attribute optimization
3. Connected belief updates to state-dependent attribute weighting
4. Defined intrinsic interpretability artifacts
5. Extended PlanningEFEMix toward explainable multi-agent planning

Questions

Appendix A: Theorem and Corollaries

Editable Theorem and Corollaries for discussion if asked.

Theorem 1	EFE minimization is structurally equivalent to MADM with state-dependent weights.
Corollary 1	Intrinsic explainability : decisions decompose into interpretable components
Corollary 2	Causal traceability : selected action can be reconstructed from the belief state and attribute contributions

Appendix B: Core Equations

Editable formulas for discussion if asked.

Variational Free Energy	$F = E\_q[\ln q(s) - \ln p(o,s)]$
Expected Free Energy	$G(\pi) = E[\ln q(s\|\pi) - \ln p(o,s\|\pi)]$
MADM-EFE Objective	$G(a) = \Sigma_k w_k(s) \cdot g_k(a)$
Softmax Policy	$\pi(a\|s) = \mathrm{softmax}(-G(a) + \mathrm{Bias}(s,a))$
Bias Update	$B_{t+1}(s,a) = \lambda B_t(s,a) + (1-\lambda)(-G(a))$

Appendix C: Theorem Proof Sketch

Structural equivalence of EFE minimization and adaptive MADM.

1. EFE decomposes into additive components $g_1 \ldots g_k$
2. Each component corresponds to an interpretable decision attribute
3. Belief updates change the relative importance of components
4. Action selection minimizes the weighted sum of components
5. The selected action can be reconstructed from attributes + weights

Therefore, the explanation is faithful because it is derived from the optimization objective itself.

Appendix D: Decision-Making Evaluation Dimensions

Evaluation should measure trade-off quality and faithfulness—not only accuracy.

Dimension	Example Metric	Interpretation
Performance	Task success, regret	Did decisions achieve outcomes?
Trade-off quality	Risk-reward balance	Were criteria balanced effectively?
Explanation faithfulness	Component reconstruction	Do explanations match the objective?
Adaptivity	Weight changes	Did priorities adjust to context?
Counterfactual clarity	Alternative-action gap	Why were rejected actions inferior?

Appendix E: Reference Anchors

Selected references supporting the framework.

- Friston (2010): Free Energy Principle
- Friston et al. (2017): Graphical brain and belief propagation
- Friston et al. (2020): Active inference on discrete state spaces
- Parr, Pezzulo & Friston (2022): Active Inference book
- Millidge, Tschantz & Buckley (2020): Expected Free Energy decomposition
- Kaelbling, Littman & Cassandra (1998): POMDP foundations
- Sutton & Barto (2018): Reinforcement learning
- Paul, Isomura & Razi (2024): Predictive planning and counterfactual learning
- Chokhawala & Mohammad (2026): PlanningEFEMix RAIS proceedings