A Proposed Study of REBUS-Informed Artificial Theory of Mind Under Partial Observability

Stephen Langsford Beale
Psychology (Psychedelics)
University of Exeter
sb1397@exeter.ac.uk

June 13, 2026

Abstract

The Inferential Robustness Improvement System (IRIS) is proposed as an evolution of a lightweight artificial Theory of Mind benchmark for testing whether improved inference produces better coordination under uncertainty. Rather than treating artificial Theory of Mind as a claim about machine consciousness or human-like mental-state understanding, IRIS operationalises the problem as context-sensitive belief updating in partially observable multi-agent environments. The proposed follow-up study tests whether a computational principle derived from REBUS — Relaxed Beliefs Under Psychedelics — can improve artificial agents' coordination by dynamically modulating the precision of prior beliefs when prediction errors increase.

REBUS, originally formulated in psychedelic neuroscience, describes how reduced precision-weighting of high-level priors may allow prediction errors to update beliefs more flexibly. This paper proposes a formal, non-phenomenological translation of that principle into machine learning: a dynamic precision or learning-rate parameter applied to a recurrent belief-state estimator. In IRIS, prediction errors such as collisions, unexpected partner switches, persistent deadlock, or high belief cross-entropy would relax prior confidence and increase belief-update sensitivity. The proposed study compares standard fixed-rate IRIS, REBUS-informed IRIS, and a combined predictive-reflective condition across two open benchmark tasks appropriated from occupational psychology evaluations.

The central hypothesis is that REBUS-informed precision modulation will improve coordination robustness by helping agents adapt when their current model of the interaction is failing. Secondary hypotheses are that such modulation may shift the existing ToMCoordScore ceiling, delay long-run degradation, and improve response to context regime changes. The study is intended as a psychology-informed AI evaluation proposal, not as a claim of equivalence between psychedelic states and artificial cognition. Its contribution is to show how a psychological theory of belief updating can be translated into a benchmarkable computational intervention.

Keywords REBUS · artificial theory of mind · partial observability · belief updating · precision modulation · multi-agent coordination

1 Hypothesis

The proposed study begins from a narrow but testable claim: artificial agents operating under partial observability may coordinate more robustly if their belief-update precision is dynamically modulated by prediction error. In standard IRIS, the agent maintains an internal belief state over partner types using a recurrent architecture. However, the update process is effectively fixed. The agent's belief-update rate does not automatically become more flexible when the policy begins to fail, when collisions increase, or when the partner's behaviour no longer matches prior expectations.

The REBUS model suggests a different principle. In biological cognition, Carhart-Harris and Friston's REBUS account proposes that psychedelics relax the precision-weighting of high-level priors, allowing prediction errors to propagate upward and update beliefs more readily. The present proposal does not claim that artificial agents possess consciousness,

subjective experience, or psychedelic-like states. Instead, it treats REBUS as a formal account of belief updating under uncertainty. The relevant mechanism is precision relaxation: when the existing model performs poorly, confidence in that model should decrease, and new evidence should have greater impact.
The main hypothesis is therefore:

H1: REBUS-informed precision modulation will improve coordination outcomes in IRIS by allowing agents to recalibrate beliefs faster when prediction errors indicate that the current interaction model is failing.
This produces several more specific predictions.
First, REBUS-informed agents should improve or shift the existing ToMCoordScore ceiling by escaping locally stable but globally suboptimal belief-action policies.
Second, precision modulation should delay or reduce long-run degradation observed in extended training, where stronger partner prediction can coincide with slower, more deadlock-prone behaviour.
Third, REBUS-informed agents should adapt more effectively to context regime changes, particularly in tasks where the same inferred partner type requires different actions under different urgency, margin, or norm conditions.
The proposed study also includes a more exploratory psychological hypothesis. In multi-agent systems, aggregate prediction-error monitoring may function as a formal analogue of detecting stale priors. When prediction errors rise across an interaction system, a REBUS-informed evaluator may infer that the current policy regime has become overconfident, brittle, or miscalibrated due to this 'cognitive load'. If so, precision modulation could be used not only inside individual agents, but also as an evaluation principle for orchestrating multi-agent systems.

2 Proposed Method

2.1 Design

The study would compare three versions of the IRIS framework across fixed benchmark tasks and fixed seed sets. The first condition would use standard IRIS, a recurrent belief-state policy with a fixed update profile. The second condition would use REBUS-informed IRIS, in which belief precision is dynamically modulated by prediction error. The third condition would use a combined predictive-reflective version, integrating standard Theory-of-Mind-style partner prediction with REBUS-informed belief flexibility and certainty-conditioned policy control.
The proposed design would use five seeds per condition at each training horizon. Each condition would be evaluated under the same benchmark harness, using the same environment, same evaluation code, same scenario suite, and same decision metrics. This is essential because the study's purpose is not merely to improve performance, but to test whether a specific psychological principle produces a measurable change in belief-action coordination.

2.2 Architecture

IRIS uses a lightweight recurrent architecture. A GRU-based belief-state estimator maintains a distribution over partner types. The model includes a belief head, partner-action prediction head, and policy head. In the standard configuration, the hidden state is updated recurrently from observations, belief logits are passed through a softmax, and the resulting belief vector is concatenated with the hidden state before action logits are generated.
The proposed REBUS modification occurs between belief computation and policy selection. In standard IRIS, belief is computed as:

```
belief = softmax(belief_logits)
```

In the REBUS-informed condition, belief precision becomes dynamic:

```
prediction_error = abs(observed_error - expected_error)
alpha_t = alpha_base + k * sigmoid(prediction_error)
belief = softmax(alpha_t * belief_logits)
```

Here, alpha_t is a time-varying precision or update-sensitivity parameter. When prediction error is low, the agent's existing belief remains relatively stable. When prediction error rises, prior confidence is relaxed and the belief distribution becomes more sensitive to new evidence. In psychological terms, the agent becomes less rigidly governed by its existing high-level prior when the environment signals that the prior is failing.
The proposal also includes certainty-conditioned policy gating. Belief entropy is already computable from the belief distribution. When entropy is high, the policy can shift away from premature commitment and toward probing or information-seeking action. When entropy is low and contextual conditions support action, the agent can commit more decisively. The key aim is to prevent two opposed failure modes: overconfident collision and passive deadlock.

2.3 Benchmarks

The proposed study would use two benchmark families.
The first is **Ambiguous Bottleneck Negotiation**. This task tests whether the agent can infer another agent's likely interaction style and coordinate through a contested passage or shared-resource bottleneck. It is suitable for testing whether REBUS modulation changes the long-run degradation boundary. Previous results suggest that extended training can improve partner inference and reduce collision while increasing delay or deadlock. The REBUS-informed condition would test whether dynamic precision relaxation maintains adaptive flexibility for longer.
The second is **Contextual Right-of-Way**. This task tests whether the same inferred partner type can lead to different actions under different contexts. Context tags include urgency, safety margin, social norm, timeout pressure, and evidence release. This task is especially well suited to the REBUS proposal because it requires agents to update not only beliefs about the partner, but also the action relevance of those beliefs. A cooperative partner may require caution under narrow safety margins; an assertive partner may require controlled assertion under high urgency. REBUS-informed flexibility should, in theory, improve adaptation to these context regime changes.

2.4 Conditions

The proposed study would compare the following conditions:

Condition	Belief update	Policy control	Predicted pattern
Standard IRIS	Fixed precision	Context-shaped recurrent policy	Strong baseline; may retain existing ToMCoordScore ceiling
REBUS-informed IRIS	Dynamic precision from prediction error	Certainty-conditioned policy	Faster recalibration after failure; possible reduction in degradation
Combined predictive-reflective IRIS	Dynamic precision plus partner prediction	Full context and certainty conditioning	Possible shift in ceiling; strongest test of belief-action integration

2.5 Outcome Measures

The primary outcome would be ToMCoordScore, a composite metric combining success, coordination efficiency, intention-prediction F1, strategy-switch accuracy, and ambiguity efficiency, minus penalties for collision, deadlock, and delay. This metric is retained because the study aims to distinguish accurate inference from effective coordination.
Secondary measures would include IntentionPredictionF1, SuccessRate, CollisionRate, DeadlockRate, AverageDelay, StrategySwitchAccuracy, AmbiguityEfficiency, and ContextSensitiveActionRegret. Belief entropy and belief cross-entropy would be treated as especially important internal diagnostics. Cross-entropy between predicted and true partner type already functions as a prediction-error signal and can therefore be reused as the basis for REBUS-informed precision modulation.

2.6 Analysis Plan

The analysis would compare each REBUS-informed condition against standard IRIS across seeds and horizons. The main test would be whether REBUS improves ToMCoordScore without merely trading collision for deadlock, or speed for unsafe commitment. A successful outcome would show improved or preserved success, reduced collision, stable or reduced deadlock, and better strategy switching under ambiguity.
The most important analysis would examine dissociations between belief and action. If IntentionPredictionF1 improves but SuccessRate falls, the intervention would have improved inference without improving coordination. If CollisionRate falls but DeadlockRate rises, the agent may have become safer but too passive. If StrategySwitchAccuracy improves alongside reduced ContextSensitiveActionRegret, this would provide stronger evidence that REBUS modulation improves the practical use of belief under changing conditions.
Long-run analysis would focus on whether REBUS shifts the apparent degradation boundary. If the standard model becomes slower or more deadlock-prone after extended training, the REBUS-informed condition should ideally maintain sensitivity to prediction error and avoid over-stabilising around brittle partner-behaviour statistics.

3 Discussion of Potential Issues and Implications

The proposed study has three main implications.

First, it offers a way to translate a psychological theory of belief updating into a machine-learning experiment without making inflated claims about artificial consciousness. REBUS is not used here as a metaphor for machine psychedelia.

It is treated as a formal mechanism: when prediction error rises, prior precision should relax and belief updating should become more flexible. This makes the psychological construct experimentally tractable.

Second, the study extends IRIS from a Theory-of-Mind benchmark into a more general Inferential Robustness Improvement System. The shift in name matters. "Theory of Mind" can imply a strong claim about mental-state attribution. IRIS instead emphasises the operational problem: how robustly does an agent update and use inferred state under uncertainty? This is a better fit for psychology-informed AI evaluation because it focuses on observable functional behaviour rather than contested claims about machine understanding.

Third, the study may clarify why improved inference sometimes fails to improve coordination. Prior IRIS-style results suggest that stronger partner prediction can reduce collisions while increasing delay, deadlock, or policy rigidity. This resembles a broader psychological problem: knowing more about a situation does not guarantee acting better within it. The agent may over-monitor, hesitate, become overconfident in a stale belief, or fail to switch when context changes. REBUS-informed precision modulation directly targets this rigidity.

Several issues would need careful handling.

The first is conceptual overreach. The biological REBUS model concerns psychedelic effects on human cognition, perception, and belief updating. A machine-learning implementation cannot reproduce the subjective or neurobiological content of that state. The study must therefore remain clear that it tests a shared formal principle, not an equivalence between humans and machines.

The second issue is parameter validity. REBUS-inspired parameters such as prediction-error sensitivity, precision relaxation, and belief-update rate may be informed by human studies, but they cannot be transferred directly without caution. The artificial benchmark has different timescales, reward structures, and error signals. Parameter sweeps would therefore be necessary. Published psychological values may guide initialisation, but the study should not assume one-to-one mapping.

The third issue is metric gaming. A precision-modulated agent might improve ToMCoordScore by becoming cautious rather than genuinely adaptive. This is why collision, deadlock, success, delay, and ambiguity efficiency must be analysed together. A safer but paralysed policy is not a robustly intelligent policy. Similarly, a faster but collision-prone policy should not be treated as successful. The value of IRIS lies in penalising both over-assertion and false safety through passivity.

The fourth issue is causal interpretation. If REBUS-informed IRIS outperforms standard IRIS, it must be shown that the improvement comes from dynamic precision modulation rather than from additional parameters, altered policy priors, or incidental regularisation. Ablation studies would be required. These should include dynamic precision without certainty-conditioned policy gating, certainty gating without dynamic precision, and full combined REBUS modulation. This would help isolate the active mechanism.

The fifth issue is ecological validity. The benchmark is intentionally small and abstract. It cannot capture the full complexity of human social cognition or psychedelic-assisted psychological change. However, its simplicity is also a strength. A small, fixed, reproducible benchmark can test a precise computational claim more cleanly than a large and opaque simulation. If REBUS-informed precision modulation fails in this setting, stronger claims would be premature. If it succeeds, the result would justify more complex follow-up studies.

The broader implication is that psychology can contribute to AI evaluation not only through human-like tasks, but through computational principles. Belief rigidity, uncertainty, prediction error, confidence, and context sensitivity are not merely humanistic descriptors. They can be operationalised as measurable properties of artificial agents. IRIS provides a candidate framework for testing such constructs in a controlled way.

4 Conclusion and Summary

This proposed follow-up study reframes the original artificial Theory of Mind benchmark as IRIS: the Inferential Robustness Improvement System. The change reflects a broader and more cautious aim. Rather than asking whether an AI system possesses Theory of Mind, IRIS asks whether an agent can update and use inferred state robustly under uncertainty.

The proposed experiment tests whether a REBUS-derived principle of precision relaxation can improve coordination in partially observable multi-agent tasks. In standard IRIS, the agent's belief update is relatively fixed. In REBUS-informed IRIS, prediction errors such as collision, unexpected partner behaviour, persistent deadlock, or high belief cross-entropy

dynamically alter the precision of belief updating. When the model's current assumptions fail, prior confidence relaxes and new evidence has greater influence.

The main hypothesis is that this mechanism will improve coordination robustness by helping agents recalibrate faster under regime change. The proposed study would compare standard IRIS, REBUS-informed IRIS, and a combined predictive-reflective condition across Ambiguous Bottleneck Negotiation and Contextual Right-of-Way tasks. The primary outcome would remain ToMCoordScore, supported by secondary measures of prediction accuracy, success, collision, deadlock, delay, strategy switching, ambiguity efficiency, and context-sensitive regret.

The study's contribution is conceptual as well as technical. It shows how a psychology-derived theory of belief updating can be translated into a formal AI evaluation intervention without claiming phenomenological equivalence between biological and artificial systems. If successful, REBUS-informed IRIS would suggest that adaptive precision modulation is a useful principle for artificial coordination under uncertainty. If unsuccessful, the result would still clarify the limits of applying psychological belief-update models to machine agents. In either case, the framework offers a disciplined route for testing how inference becomes action, and when it fails to do so.

References

[1] Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. Pharmacological reviews, 71(3), 316–344. DOI: 10.1017/s0033291715002901

[2] Caulfield, A., Young, A. H., & Mehta, M. (2025). Learning as the unifying mechanism of psychedelic action. Journal of psychopharmacology (Oxford), 2698811251405683. https://doi.org/10.1177/02698811251405683

[3] Caulfield, A., Li, L., Askari, F., Belessiotis-Richards, C., Moura, R., Young, A., & Mehta, M. (2026). The effect of psychedelics on associative learning: a systematic review. Neuroscience Applied, 5, 106386. https://doi.org/10.1016/j.nsa.2025.106386

[4] Kanen, J. W., Luo, Q., Rostami Kandroodi, M., Cardinal, R. N., Robbins, T. W., Nutt, D. J., Carhart-Harris, R. L., & den Ouden, H. E. M. (2023). Effect of lysergic acid diethylamide (LSD) on reinforcement learning in humans. Psychological Medicine, 53(14), 6434–6445. https://doi.org/10.1017/S0033291722002963

[5] Kanen, J., Luppi, A., Luo, Q., Roseman, L., Cardinal, R., Robbins, T., Nutt, D., Carhart-Harris, R., & den Ouden, H. (2025). Neural and behavioral evidence from reinforcement learning converge to support the REBUS (RElaxed Beliefs Under pSychedelics) model under LSD. The international journal of neuropsychopharmacology, 28(Supplement_2), ii82–ii82. https://doi.org/10.1093/ijnp/pyaf052.163