FROM INFERENCE TO COORDINATION: A LIGHTWEIGHT BENCHMARK FOR CONTEXT-SENSITIVE ARTIFICIAL THEORY OF MIND UNDER PARTIAL OBSERVABILITY

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

June 13, 2026

ABSTRACT

Claims about artificial Theory of Mind are often assessed through prediction tasks: whether a model can infer another agent's belief, intention, or likely next action. However, successful prediction is not the same as successful coordination. This paper presents a lightweight two-agent benchmark for testing whether improved partner inference produces better action under partial observability. The benchmark models an ambiguous right-of-way negotiation in which an agent must decide when to proceed, wait, probe, or yield while observing only partial evidence of the other agent's latent interaction style. The task is framed not as proof of machine mind-reading, but as a pragmatic test of context-sensitive coordination under uncertainty.

A recurrent belief-state policy was trained and evaluated against a fixed benchmark harness. The model used auxiliary opponent-state prediction and context-conditioned policy shaping, with explicit contextual variables including urgency, norm, safety margin, and evidence release. The primary composite metric, ToMCoordScore, combines success, coordination efficiency, intention-prediction F1, strategy-switch accuracy, ambiguity efficiency, and penalties for collision, deadlock, and delay.

Across five seeds at 800 training episodes, the candidate model improved mean ToMCoordScore from 0.1109 to 0.2054, increased intention-prediction F1 from 0.0000 to 0.4141, reduced collision rate from 0.34 to 0.32, and slightly increased success rate from 0.56 to 0.58. At 140,000 episodes, ToMCoordScore rose further to 0.3828 and collision rate fell to 0.11, but deadlock increased to 0.13 and success fell to 0.56. These findings suggest that better inference can improve coordination, but also reveal a "hard pocket" in which accurate belief and correct action come into tension under operational pressure. The benchmark therefore offers a compact method for studying the inference-to-coordination gap in artificial agents.

Keywords artificial theory of mind · partial observability · coordination · opponent modelling · belief-state learning · multi-agent benchmarks

1 Introduction and Literature Review

Research on artificial Theory of Mind has expanded rapidly as AI systems have become increasingly agentic, interactive, and embedded in multi-agent settings. In psychology, Theory of Mind refers to the capacity to attribute beliefs, intentions, desires, and perspectives to others. In machine learning, however, the term is often operationalised more narrowly: a model is said to display Theory-of-Mind-like behaviour if it predicts another agent's belief state, intention, or likely next action. This translation is useful, but incomplete. A system may predict another agent accurately while still acting poorly in interaction. The present study therefore treats artificial Theory of Mind not as a claim about machine mentality, but as a functional question: does inference about another agent improve coordination under uncertainty?

The distinction is important because prediction and coordination impose different demands. Prediction asks whether

an agent can classify another agent's type or likely behaviour. Coordination asks whether that classification supports timely, context-sensitive action. In human social cognition, the adaptive value of Theory of Mind does not lie merely in identifying another person's state, but in using that inference to regulate one's own behaviour: when to speak, wait, challenge, defer, reassure, interrupt, or act. The same distinction applies to artificial agents. A system that identifies an assertive partner but yields indefinitely has inferred something useful but failed to coordinate. Conversely, a system that identifies a cooperative partner but proceeds under unsafe conditions may translate belief into premature commitment.
This problem is especially relevant in partially observable environments. Partially Observable Markov Decision Processes provide a formal account of decision-making where the true state of the system cannot be directly observed. In such settings, an agent must maintain an approximate belief state over hidden variables and map that belief to action. Classical POMDP methods are powerful but computationally difficult: belief spaces are continuous, histories are large, and exact solutions often scale poorly. Applied work therefore commonly relies on approximate, heuristic, or learned representations of belief. The present benchmark follows this practical direction by using a lightweight recurrent state estimator rather than a full symbolic POMDP solver.
The benchmark also draws from opponent modelling and belief-state learning. In opponent modelling, an agent estimates the likely behaviour or latent type of another agent. Related work on social reasoning in AI, including Bayesian and neural approaches to Theory of Mind, has shown that models can learn useful representations of other agents. However, such work can invite over-mentalistic interpretation if predictive success is treated as evidence of genuine social understanding. The current study deliberately avoids this stronger claim. Its aim is narrower: to test whether a compact model can learn control-relevant summaries of partner behaviour and use them to coordinate more effectively.
The psychological motivation is therefore pragmatic. Many real-world interactions involve ambiguity, time pressure, social norms, and incomplete evidence. Competent action requires not only inference, but also judgement about the local meaning of that inference. Under one context, an assertive partner may require yielding; under another, yielding may create delay, deadlock, or unsafe over-deference. The benchmark operationalises this problem through an ambiguous right-of-way task where the same inferred partner type can require different actions depending on urgency, margin, norm, and progress. This allows artificial Theory of Mind to be evaluated not merely as prediction, but as situated coordination.

2 Method

The study used a fixed two-agent negotiation benchmark under partial observability. Agent A and Agent B approach a contested right-of-way decision, modelled as a small shared-resource coordination problem. Only one agent can proceed cleanly at a time. Agent A observes local state, recent behaviour from Agent B, and explicit context tags, but does not observe Agent B's latent partner style, short-horizon goal, or whether B will maintain or reverse its current posture. The central task is for Agent A to decide whether to proceed, wait, probe, or yield.
The benchmark includes hidden partner styles such as cooperative, assertive, hesitant, opportunistic, and deceptive-switching. It also includes contextual variables that alter the appropriate action: urgency, safety margin, social norm, timeout pressure, and evidence release. These variables are essential to the design. The benchmark is not intended to reward a static mapping such as "cooperative partner means proceed" or "assertive partner means yield". Instead, it tests whether the agent can integrate inferred partner state with explicit contextual demands. A cooperative partner may still require caution under narrow safety margins; an assertive partner may still need to be challenged under urgency or throughput pressure.
The active model was a lightweight recurrent policy with an internal belief-state representation. The policy used a recurrent estimator, auxiliary opponent-state prediction, and context-conditioned action priors. Training was bounded to a single editable surface, train.py, while the benchmark environment and evaluation code were fixed to preserve comparability. This design follows an autoresearch-style loop: constrained code changes, fixed-budget experiments, stable metrics, and promotion or rejection of candidate changes according to explicit decision criteria. The purpose was not to use autoresearch as a POMDP solver, but to borrow its disciplined experimental pattern.
Candidate models were first evaluated through a 1/3/5 ladder: a one-seed smoke test, a three-seed quick gate, and a five-seed promotion gate. The scientific promotion gate used seeds 7, 11, 17, 23, and 29, with 800 training episodes per seed. A candidate was retained only if mean ToMCoordScore improved, mean deadlock did not worsen, collision rate was lower or equal, success rate was higher or equal, and there was no catastrophic single-seed regression. Longer inference evaluations were then conducted at 140,000 episodes to examine whether improved belief learning continued to translate into improved coordination.
The primary outcome was ToMCoordScore, a composite score designed to distinguish better belief, better action, and false safety through passivity. It combines SuccessRate, CoordinationEfficiency, IntentionPredictionF1, StrategySwitchAccuracy, and AmbiguityEfficiency, then subtracts penalties for CollisionRate, DeadlockRate, and AverageDelay. Secondary diagnostics included intention-prediction F1, collision rate, deadlock rate, success rate, strategy-switch

accuracy, ambiguity efficiency, and average delay. Motivation-aware logs classified actions by context and action style, including high-conflict bottleneck, late-phase, normal-flow, self-priority, cooperative-yield, and cooperative-wait categories.

3 Results

3.1 Five-seed comparison at 800 training episodes

At 800 episodes, the candidate model outperformed the baseline on the main composite measure. Mean ToMCoordScore increased from 0.1109 to 0.2054. IntentionPredictionF1 increased from 0.0000 to 0.4141, indicating that the candidate learned a meaningful diagnostic representation of partner type. CollisionRate fell from 0.34 to 0.32, while SuccessRate rose from 0.56 to 0.58. DeadlockRate remained unchanged at 0.10.

Metric	Baseline mean	Candidate mean	Change
ToMCoordScore	0.1109	0.2054	+0.0945
DeadlockRate	0.1000	0.1000	0.0000
CollisionRate	0.3400	0.3200	-0.0200
SuccessRate	0.5600	0.5800	+0.0200
AmbiguityEfficiency	0.0900	0.1267	+0.0367
IntentionPredictionF1	0.0000	0.4141	+0.4141
StrategySwitchAccuracy	0.5000	0.5400	+0.0400

The result satisfied the promotion criterion because the candidate improved the main composite score, did not worsen mean deadlock, did not increase collision rate, and did not reduce success rate. The improvement in F1 was therefore not isolated from action quality; it was accompanied by modest but meaningful gains in coordination outcomes.

3.2 Long-run evaluation at 140,000 episodes

At 140,000 episodes, the candidate showed much stronger inference and reduced collision, but also introduced new coordination costs. ToMCoordScore increased from the 800-episode candidate level of 0.2265 to 0.3828. IntentionPredictionF1 increased from 0.4148 to 0.6989. CollisionRate fell sharply from 0.2900 to 0.1100. However, DeadlockRate increased from 0.1000 to 0.1300, SuccessRate fell from 0.6100 to 0.5600, and AverageDelay increased from 11.95 to 15.29.

Metric	Candidate at 800 episodes	Candidate at 140k episodes	Change
ToMCoordScore	0.2265	0.3828	+0.1563
DeadlockRate	0.1000	0.1300	+0.0300
CollisionRate	0.2900	0.1100	-0.1800
SuccessRate	0.6100	0.5600	-0.0500
IntentionPredictionF1	0.4148	0.6989	+0.2841
AverageDelay	11.95	15.29	+3.34

Per-seed results showed consistent improvement in ToMCoordScore but mixed effects on clean task resolution. Seed 29 achieved the highest ToMCoordScore at 0.4152, with SuccessRate of 0.60 and CollisionRate of 0.05. Seed 11 was weakest, with ToMCoordScore of 0.3434, SuccessRate of 0.50, and DeadlockRate of 0.15. The long-run pattern therefore suggests that the model learned better belief and better collision avoidance, but did not universally convert these gains into higher success.

Seed	Score	Success	Collision	Deadlock	F1	Switch acc.
29	0.4152	0.60	0.05	0.15	0.6923	0.90
7	0.3921	0.60	0.10	0.10	0.6863	0.85
23	0.3895	0.55	0.10	0.15	0.7142	0.90
17	0.3736	0.55	0.15	0.10	0.6885	0.95
11	0.3434	0.50	0.15	0.15	0.7129	0.90

3.3 Analysis

The central result is that improved inference did improve coordination, but only partially. The 800-episode model showed the cleanest form of the intended effect: partner-state prediction rose sharply, ToMCoordScore improved,

deadlock did not worsen, collision fell slightly, and success rose slightly. This suggests that the benchmark can identify cases where belief-state learning becomes control-relevant rather than merely diagnostic.

The 140,000-episode model produced a more complex pattern. On the positive side, the model became much better at partner-type prediction and much less collision-prone. This indicates stronger belief formation and more cautious calibration around commitment. It also improved StrategySwitchAccuracy, suggesting better sensitivity to when a policy should move from caution to action or from assertion to yielding. These are meaningful coordination gains.

However, the same long-run model became slower and somewhat more deadlock-prone, and its success rate fell relative to the 800-episode candidate. This pattern is psychologically and computationally important. It suggests that belief quality alone is not sufficient for competent interaction. A model can become better at reading the situation while also becoming more conservative, stickier, or less efficient in converting inference into action. In human terms, this resembles over-monitoring or excessive deliberation: the agent knows more, but acts less cleanly.

This is the basis for the "hard pocket" interpretation. A hard pocket is a bounded region of the coordination space in which accurate inference and appropriate action come into tension. Under high urgency, narrow margins, throughput-biased norms, or conflicting partner signals, the agent may correctly identify the partner's posture but still make a regrettable trade-off. It may yield when urgency requires assertion, assert when margin requires caution, or wait when the evidence is already sufficient to proceed. The problem is not simple ignorance. It is miscalibrated use of knowledge under pressure.

The benchmark therefore separates three forms of performance that are often conflated: inference, action selection, and coordination outcome. IntentionPredictionF1 measures whether the agent identifies the other agent. StrategySwitchAccuracy and AmbiguityEfficiency measure whether the agent changes behaviour appropriately. Success, collision, deadlock, and delay measure whether the interaction actually resolves well. The divergence between these metrics at 140,000 episodes is the study's most important finding. It shows why an artificial Theory of Mind benchmark should not stop at prediction.

The result also supports the value of lightweight architectures. Because the model is small and the benchmark is fixed, the observed ceiling is unlikely to be hidden behind large-scale capacity effects. The limits are more plausibly structural, informational, or policy-calibration related. This makes the benchmark useful not only for demonstrating improvement, but for locating the conditions under which improvement fails.

4 Discussion

The present study contributes a compact benchmark for evaluating artificial Theory-of-Mind-like behaviour as applied coordination rather than isolated prediction. The findings support the claim that recurrent belief-state learning and auxiliary opponent modelling can improve coordination under partial observability. However, they also show that the relationship between inference and coordination is not linear. Better prediction may reduce some harms, especially collision, while increasing other costs, such as delay or deadlock.

This distinction has psychological significance. Human Theory of Mind is not merely a representational capacity; it is embedded in regulation, timing, inhibition, confidence, and social norm use. A person who correctly infers another's intention may still act poorly if the situation demands speed, restraint, challenge, or tact. The same appears true in simplified artificial agents. The benchmark's results suggest that the applied value of "mind-reading" depends on the action policy that uses it. In this sense, artificial Theory of Mind should be evaluated not only by whether a system forms an accurate belief, but by whether that belief improves situated conduct.

The hard pocket finding is particularly important. In many AI evaluations, residual failure is treated as noise, insufficient training, or an engineering defect. Here, failure is more informative. The hard pocket identifies a regime in which the agent's internal representation is not simply wrong, but insufficiently actionable. High urgency, narrow safety margins, conflicting norms, and partial evidence create conditions where the correct action may oppose the most obvious inference. For example, a cooperative partner may invite proceeding, but a narrow margin may require caution. An assertive partner may invite yielding, but time pressure may require probing or controlled assertion. These cases cannot be solved by inference alone.

This has implications for AI deployment. Human-facing assistants, multi-agent software systems, robotics, traffic coordination, scheduling systems, and distributed service agents all face situations where unilateral optimisation can create friction or breakdown. The practical challenge is often not whether an agent can act, but whether it should act now, wait, request clarification, yield, interrupt, or escalate. Systems that perform well in ordinary conditions may degrade under time pressure, conflicting authority, or incomplete information. A benchmark that explicitly reports such limits is therefore more useful than one that reports only aggregate success.

The results also caution against over-claiming artificial Theory of Mind. The model does not demonstrate human-like mental-state understanding. It learns a compact control-relevant representation of partner behaviour and combines that representation with explicit context tags. That is still valuable. Indeed, the more modest framing may be scientifically stronger. The relevant question is not whether the agent "has" Theory of Mind, but whether a belief-like internal state

improves coordination when the environment is uncertain and socially contingent.
There are several limitations. First, the benchmark is intentionally small. It uses a simplified two-agent right-of-way structure rather than a rich human social environment. Second, partner types are predefined and limited. Third, context tags such as urgency and margin are mostly explicit rather than inferred, which reduces ecological complexity. Fourth, the current model includes hand-engineered context-conditioned policy shaping, making it difficult to fully separate learned belief from designed priors. Fifth, the long-run findings are seed-sensitive, suggesting that training horizon and stopping criteria require further study.
Future work should therefore separate belief learning and policy control more cleanly. This could include comparing a plain recurrent baseline, a belief-only model, a context-only model, and a fused belief-context model. It would also be useful to move from static partner type to dynamic latent intention, where the other agent can change goals or posture during the episode. Generalisation should be tested on held-out partner policies and scenario compositions. Finally, the hard pocket should be studied directly: whether it is a property of this architecture, this training regime, or a more general information-theoretic limit on coordination under partial observability.
Despite these limitations, the study demonstrates a practical route for evaluating socially relevant agent behaviour without frontier-scale compute. The model is small, the benchmark is reproducible, and the metrics distinguish prediction from action. This makes the framework suitable for iterative research into applied artificial social cognition, especially where psychology-informed concepts need to be translated into operational machine-learning tests.

5 Conclusion and Summary

This paper presented a lightweight benchmark for evaluating whether artificial Theory-of-Mind-like inference improves coordination under partial observability. Rather than treating Theory of Mind as literal machine mind-reading, the study operationalised it as a recurrent belief-state estimate of another agent's latent interaction style, combined with context-sensitive action selection. The benchmark used a two-agent ambiguous right-of-way task in which the correct action depended not only on inferred partner type, but also on urgency, safety margin, norm, timeout pressure, and evidence release.
The results show that improved inference can produce better coordination. At 800 training episodes, the candidate model improved ToMCoordScore, increased intention-prediction F1, reduced collision rate, slightly improved success rate, and did not increase deadlock. This supports the hypothesis that lightweight belief-state learning can become control-relevant in a socially contingent coordination task.
Long-run evaluation at 140,000 episodes produced a more complex result. The model achieved much stronger intention prediction and substantially reduced collision, but also became slower, somewhat more deadlock-prone, and less successful overall than the 800-episode candidate. This divergence is the study's main theoretical contribution. It shows that better belief does not automatically imply better action. In some regions of the task space, accurate inference and appropriate conduct come into tension. The paper describes this as the hard pocket problem: a bounded regime where context, pressure, and belief conflict.
The benchmark therefore offers a method for studying the inference-to-coordination gap. Its value lies not in proving that artificial agents possess human-like Theory of Mind, but in measuring when belief-like representations improve action, when they fail, and what kinds of context make the difference. This makes the framework relevant to psychology-informed AI evaluation, multi-agent coordination, and the design of agents that must act tactfully under uncertainty.

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.

[2] Chades, I., Pascal, L. V., Nicol, S., Fletcher, C. S., & Ferrer-Mestres, J. (2021). A primer on partially observable Markov decision processes (POMDPs). Methods in Ecology and Evolution, 12(11), 2058–2072. https://besjournals.onlinelibrary.wiley.com/doi/epdf/10.1111/2041-210X.13692

[3] Cuzzolin F, Morelli A, Cîrstea B, Sahakian BJ (2020). Knowing me, knowing you: theory of mind in AI. Psychological Medicine 1 5. https://doi.org/10.1017/S0033291720000835

[4] Hoerger, M., Kurniawati, H., & Elfes, A. (2023). Multilevel Monte Carlo for solving POMDPs on-line. The International Journal of Robotics Research, 42(4-5), 196–213. https://link.springer.com/chapter/10.1007/978-3-030-95459-8_11

[5] Hoerger, M., Kurniawati, H., Kroese, D., & Ye, N. (2022). Adaptive discretization using voronoi trees for continuous-action pomdps. International Workshop on the Algorithmic Foundations of Robotics. https://arxiv.org/abs/2107.07599

[6] Krasnytskyi, N., & Cuzzolin, F. (2025). Evaluating Machine Theory of Mind: A Critical Analysis of ToMnet-N. ToM4AI 2025, 4, 60. https://www.academia.edu/download/123035967/Proceedings_of_1st_Workshop_on_Advancing_Artificia.pdf#page=63

[7] Kurniawati, H. (2022). Partially observable markov decision processes and robotics. Annual review of control, robotics, and autonomous systems, 5(1), 253–277. https://arxiv.org/abs/2107.07599

[8] Lebiere, C., Pirolli, P., Johnson, M., Martin, M., & Morrison, D. (2025). Cognitive Models for Machine Theory of Mind. Topics in Cognitive Science, 17(2), 268–290. https://doi.org/https://doi.org/10.1111/tops.12773

[9] Lee, T., & Kim, Y. J. (2013). GPU-based motion planning under uncertainties using POMDP. 2013 IEEE international conference on robotics and automation. https://ieeexplore.ieee.org/document/6631227

[10] Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. International conference on machine learning. https://proceedings.mlr.press/v80/rabinowitz18a.html

[11] Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A., Loftus, A. R., Jannali, A. R., Prakash, N., Cui, J., Rogers, G., Brinkmann, J., Rager, C., Zur, A., Ripa, M. U., Sankaranarayanan, A., Atkinson, D., Gandikota, R.,. . . Bau, D. (2026). Agents of Chaos. https://arxiv.org/abs/2602.20021

[12] Spaan, M. T. (2012). Partially observable Markov decision processes. Reinforcement learning: State-of-the-art, 387–414. https://link.springer.com/chapter/10.1007/978-3-642-27645-3_12

[13] Sultana, M., Yorke-Smith, N., Wang, K., Manchingal, S. K., Mubashar, M., & Cuzzolin, F. (2025). Epistemic wrapping for uncertainty quantification. arXiv preprint arXiv:2505.02277. https://arxiv.org/abs/2505.02277

[14] Wang, et al. (2023). Believer: Belief-state modelling for partially observable reinforcement learning.

Appendix A: Ambiguous Bottleneck Task

The Ambiguous Bottleneck task is a compact two-agent coordination problem. Two agents approach a constrained passage, merge point, or shared-resource gate. Only one can pass cleanly at a time. The focal agent must choose whether to proceed, wait, yield, or probe while observing incomplete evidence about the other agent's likely behaviour. The task is designed to create ambiguity between cooperation and conflict. The other agent may appear cooperative, assertive, hesitant, opportunistic, or inconsistent. Its latent type is not directly visible. The focal agent must infer this type from recent behaviour, while also considering whether the current phase of the episode supports decisive action or continued caution.

The task penalises both premature commitment and excessive passivity. A collision indicates that the agent acted too assertively or misread the interaction. A deadlock indicates that the agent waited or yielded too long. A timeout indicates failure to convert available evidence into resolution. Success requires the agent to balance caution, inference, and timely switching.

The Ambiguous Bottleneck task is therefore not merely a test of partner classification. It tests whether inferred belief is translated into better action under conditions where delay, clash, and hesitation all carry costs.

Appendix B: Contextual Right-of-Way Task

The Contextual Right-of-Way task extends the bottleneck design by making context central. In this variant, the same inferred partner type can require different actions depending on the situation. A cooperative partner may not always justify proceeding; an assertive partner may not always justify yielding. The correct action depends on the conjunction of inferred partner style and explicit context tags.

The task includes context variables such as urgency, safety margin, social norm, timeout pressure, and evidence release. Scenario families include same-belief-different-action, urgency override, safety first, opportunism under norm shift, and social misread recovery. These scenario families are designed to prevent brittle static mappings between partner type and action.

The key psychological idea is that social competence is not simply knowing "who the other is". It is knowing what that information means in context. The Contextual Right-of-Way task therefore evaluates whether an agent can switch

between restraint, assertion, probing, and yielding as contextual demands change.
The most important failure modes include correct partner read but wrong contextual action, excessive politeness under urgency, justified caution becoming deadlock, static social mapping, and late use of evidence. For each run, the benchmark logs scenario family, partner style, context tag set, belief turning point, action switch point, outcome, and interpretation.