INTENTION INFERENCE PRODUCES THE RETALIATORY CASCADES IT PREVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In noisy, repeated games, inferring partner intent can explain an observed defection as an execution error. We study how this ability affects learning and cooperation in Bayesian model-based learners using active inference (AIF) and approximate Bayes-adaptive planning. For symmetric 2×2 games, we derive how execution noise changes payoff comparisons and the information available about latent intent, and formulate decision boundaries for Bayesian model-based planners. When contrasting these generative model architectures, intent-conditioned learners lose cooperation sharply in Stag Hunt and the iterated Prisoner's Dilemma (IPD). Against an unconditional cooperator in Stag Hunt, raising noise from 0.15 to 0.20 reduces mean late intended cooperation from 0.927 to 0.201, compared with 0.887 to 0.695 for the observation-conditioned learner, although cooperation remains the better one-round response to the known partner. AIF learning traces show that runs with different eventual outcomes can have similar beliefs about current partner intent but different predictions about whether mutual cooperation will continue. We hypothesise that defection helps preserve unfavourable predictions by reducing opportunities to test continued cooperation. Observation-conditioned clusters are well described by learned responses to executed outcomes; some intent-conditioned AIF runs instead alternate rapidly between C and D. Bayes-adaptive mutual cooperation is lower in IPD self-play than against Tit-for-Tat, and mainly gives way to asymmetric play in Chicken. Both intent-conditioned planners retain high intended agreement in Pure Coordination self-play. These findings motivate reporting per-run behaviour and intended and executed joint actions alongside mean cooperation. They also motivate research into how inferred intent could guide exploration toward poorly learned transitions and stabilise learning under execution noise.

1 INTRODUCTION

The emergence of cooperation under uncertainty is a foundational problem across many domains, from economics to biology to artificial intelligence (Dafoe et al., 2020). Reaching a jointly preferred outcome can be straightforward when agents observe each other's actions cleanly; it may become fragile the moment actions or observations are corrupted. Execution noise is a common feature of deployed systems, making coordination through imperfect action channels a practical concern.

In a noisy repeated game, an observed non-cooperative action could be deliberate hostility or cooperative intent corrupted by noise. Strategies that treat accidental defection as intended can over-retaliate and trigger punishment cascades (Molander, 1985; Wu & Axelrod, 1995; Au & Nau, 2006; Hutter, 2023). A natural Bayesian approach is intent inference: treat partner intent as a latent variable, compute a posterior over it, and condition the policy on the posterior rather than the raw observation (Baker et al., 2009; Foerster et al., 2019). A known channel connects latent intent to its observable consequences. In repeated social dilemmas, this allows an agent to attribute accidental defections to noise and potentially avoid retaliation.

We study Bayesian model-based learners using active inference (AIF) and approximate Bayes-adaptive planning. When contrasting these generative model formulations, we find that intent-conditioned learners lose cooperation sharply, while observation-conditioned learners show a more gradual decline (Figure 1). Against an unconditional cooperator in Stag Hunt, this loss occurs while cooperation remains the strictly higher-payoff choice.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

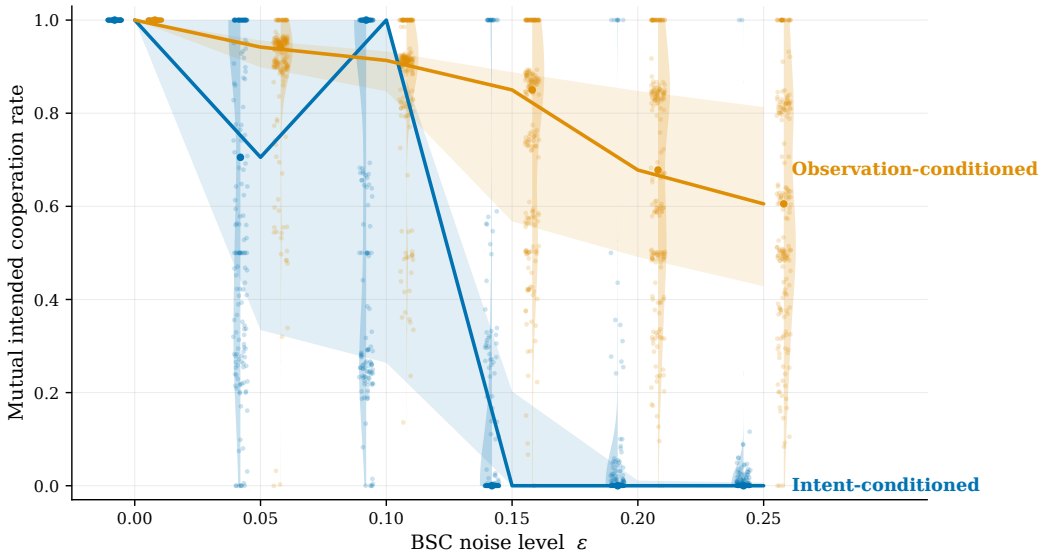


Figure 1: **Sharp cooperation loss with intent-conditioned learning.** Self-play on the canonical Prisoner's Dilemma, comparing intent-conditioned active inference (blue) with observation-conditioned active inference (orange). As noise increases, the intent-conditioned agent develops a bimodal seed distribution and loses cooperation sharply, whereas the observation-conditioned agent shows a more gradual population-level decline. Points show intended mutual cooperation for individual seeds across the final 25% of the trajectory (200 seeds, 3000 rounds); lines and bands summarise the median and interquartile range.

We model execution as a binary symmetric channel (Nowak & Sigmund, 1993; Wu & Axelrod, 1995; Molander, 1985) to connect payoff boundaries with the information available about intent. In the AIF learning traces, eventual cooperators and defectors show similar confidence in current partner cooperation intent but differ in their predictions of continued mutual cooperation. We hypothesise that non-cooperative sequences can preserve pessimistic predictions by reducing opportunities to test continued cooperation.

Per-run trajectories distinguish learned responses to executed outcomes from rapid C–D alternation; cross-game comparisons show how partner responses and payoffs change the consequences of cooperation loss.

Contributions. (1) We derive noise-adjusted payoff boundaries and quantify information about latent intent in symmetric 2×2 games (§3). (2) We formulate a planner's cooperation margin and a decision-error probability for independent observations in a fixed context (§4, Appendix C). (3) We relate the matched AIF cooperation contrast to learned predictions, conditional response rules, and observed action cycles (§5). (4) We compare AIF and approximate Bayes-adaptive planning across four games and two partner conditions (§6).

2 Related Work

Cooperation under execution noise. Reciprocity can sustain cooperation in iterated games (Axelrod & Hamilton, 1981), but execution errors can destabilise it (Molander, 1985). Proposed remedies include probabilistic forgiveness (Glynatsi et al., 2024), internal standing variables distinguishing execution from perception errors (Boerlijst et al., 1997), pattern matching over accident versus intention (Au & Nau, 2006), and memory-n strategies combining cooperation, error correction, and retaliation (Hilbe et al., 2017). Folk-theorem extensions characterise sustainable payoffs under imperfect monitoring (Fudenberg et al., 1994). We study how cooperation develops when agents learn opponent dynamics online while inferring intent through a known execution channel.

Opponent modelling and intent inference. Interactive POMDPs embed recursive opponent models into the state space (Gmytrasiewicz & Doshi, 2005); we use a non-recursive model with a Dirichlet posterior over reactive opponent dynamics. Bayesian Theory of Mind (Baker et al., 2009) and neural meta-learning approaches (Rabinowitz et al., 2018) treat opponent mental states as hidden causes of observed behaviour. The Bayesian Action Decoder factors public belief from private observations in cooperative games (Foerster et al., 2019), while differentiable opponent models account for co-adaptation (Foerster et al., 2018). Our comparison focuses on how separating execution noise from intent affects the conditional predictions learned from experience.

Bayesian learning and planning. Bayes-adaptive POMDPs plan through posterior uncertainty over transition dynamics (Pineau et al., 2008). Analytical MDP models have derived phase-transition conditions for cooperation in the IPD without execution noise (Demekas et al., 2023), while factorised active-inference models represent opponents' internal states in iterated general-sum games (Ruiz-Serra et al., 2024). In this work, we connect noise-adjusted decision boundaries with the information available about intent and examine learned behaviour under AIF and approximate Bayes-adaptive planning. The analysis distinguishes the channel's effect on estimation from the action-dependent evidence used to learn cooperative responses.

3 The Belief-Space Geometry of Noisy 2×2 Games

3.1 Symmetric 2×2 matrix games

We consider symmetric two-player matrix games with action set $\{C, D\}$, where C denotes cooperation and D defection, and payoffs (R, T, P, S), where R is mutual cooperation, P mutual defection, T unilateral defection (temptation), and S unilateral cooperation (sucker). Examples of payoff orderings include the Prisoner's Dilemma ($T > R > P > S$), Chicken ($T > R > S > P$), and the Stag Hunt ($R > T \geq P > S$). Players repeat the stage game.

The reward function is defined on *executed* joint actions, $r : \mathcal{O} \to \mathbb{R}$, where $\mathcal{O} = \{CC, CD, DC, DD\}$ and $r(o)$ is the learner's payoff for outcome o, with its own action listed first. Payoffs depend on what was played, not on what was intended, where intent denotes the action selected before execution noise. Execution noise creates a gap between the intended action and its observed, payoff-generating consequences; this gap creates the attribution problem studied here.

3.2 Cooperative boundary in the noiseless game

For the Stag Hunt, an agent with belief $b = \Pr(\text{partner intends } C)$ myopically prefers to cooperate iff $bR + (1-b)S > bT + (1-b)P$, equivalently

$$b \;>\; b^\star \;=\; \frac{P - S}{(R - T) + (P - S)}. \tag{1}$$

This payoff-defined threshold separates two one-round best-response regions: cooperate above $b^\star$ and defect below. It also characterises risk dominance: the cooperative equilibrium is risk-dominant when $b^\star < 1/2$ (Harsanyi & Selten, 1988). For example, the Stag Hunt $(R, T, P, S) = (3, 2, 1, 0)$ has $b^\star = 1/2$ (Skyrms, 2001). Both $R - T$ and $P - S$ are positive in the Stag Hunt game, so $b^\star \in (0, 1)$; for the Prisoner's Dilemma, no cooperative best-response threshold exists, although a multistep threshold can exist against reactive partners (§6).

3.3 Execution noise and the attribution problem

We model execution noise as a binary symmetric channel (BSC) with crossover probability $\epsilon \in [0, 0.5)$: each player's intended action is independently flipped with probability ϵ before realisation. The agent observes only the executed outcome, not the partner's intent.

Execution noise creates an attribution problem at the level of single observations. For a fixed context s, let the partner's probability of intending to cooperate be $\theta = \Pr(\text{partner intends } C \mid s)$. The binary symmetric channel (BSC) transforms this into an expected observable cooperation rate of

$$\phi(\theta, \epsilon) \;=\; (1 - 2\epsilon)\,\theta + \epsilon. \tag{2}$$

This mapping contracts the latent intent probability toward $1/2$ by a factor of $(1-2\epsilon)$. Because the channel can flip any action before realisation, distinct intentions can produce identical observed outcomes, forcing the learner to infer true intent from ambiguous evidence.

Execution noise deforms the intent-space boundary of equation 1. Let $\widetilde{R}, \widetilde{T}, \widetilde{P}, \widetilde{S}$ be expected payoffs conditional on each intended joint action, averaging over both players' execution flips. Their differences are

$$\widetilde{R} - \widetilde{T} = (1-2\epsilon)\big[(1-\epsilon)(R-T) - \epsilon(P-S)\big],$$
$$\widetilde{P} - \widetilde{S} = (1-2\epsilon)\big[(1-\epsilon)(P-S) - \epsilon(R-T)\big],$$

and the corresponding noise-corrected threshold is

$$b^\star_\epsilon \;=\; \frac{(1-\epsilon)(P-S) - \epsilon(R-T)}{(1-2\epsilon)\left[(R-T) + (P-S)\right]}. \tag{3}$$

When $R-T = P-S$, as in the preceding example, the corrections cancel and $b^\star_\epsilon = b^\star$. Otherwise the threshold drifts monotonically, contracting or expanding the cooperative best-response region in intent space. Equivalently, $b^\star_\epsilon = (b^\star - \epsilon)/(1-2\epsilon)$, so its representation is $\phi^\star(\epsilon) = \epsilon + (1-2\epsilon)b^\star_\epsilon = b^\star$. The observable-cooperation cutoff remains fixed in this one-round comparison, while attainable observation rates contract to $[\epsilon, 1-\epsilon]$. When $b^\star_\epsilon \geq 1$, cooperation is no longer a strict best response even to an always-cooperating partner.

3.4 Intent modelling: factoring noise from dynamics

The architectural choice is what to treat as state. The intent-conditioned learner represents latent intentions, separating the known execution channel from learned opponent dynamics. The observation-conditioned learner represents executed outcomes and learns their dynamics directly. Both induce decision boundaries.

For equivalent predictions, $\hat{\phi} = \epsilon + (1-2\epsilon)\hat{\theta}$, the tests $\hat{\theta} > b^\star_\epsilon$ and $\hat{\phi} > b^\star$ give the same decision. Differences depend on learning and planning: the intent-conditioned model informs latent-state inference, while inferred states determine the evidence updating that model (§4).

Section 4.3 extends this comparison to planners that value future rewards and information.

3.5 The information cost of intent recovery

For a fixed context, known noise rate, and t independent BSC observations, the Fisher information about intent is

$$I_t(\theta) \;=\; \frac{t\,(1-2\epsilon)^2}{\phi(\theta,\epsilon)\big(1-\phi(\theta,\epsilon)\big)}, \tag{4}$$

Under continuity conditions, its inverse lower-bounds the variance of an unbiased estimator. As $\epsilon \to 1/2$, observations become uninformative about intent and this bound grows. Bayesian posterior variance remains bounded and depends on the prior and evidence.

Transforming between observable cooperation ϕ and latent intent θ rescales posterior width and distance to the boundary equally (Figure 2). The wider intent posterior alone therefore leaves the matched one-round decision unchanged.

The information calculation above concerns observations of one fixed context. During interaction, actions determine which contexts supply evidence, while latent-state inference determines how that evidence is assigned to the conditional predictions used in planning.

4 Methodology

Having established how execution noise distorts the information available about partner intent, we now define the generative models and planning algorithms used to investigate its behavioural consequences. Our goal is to isolate how separating the execution channel from learned dynamics affects an agent's susceptibility to cooperation collapse.

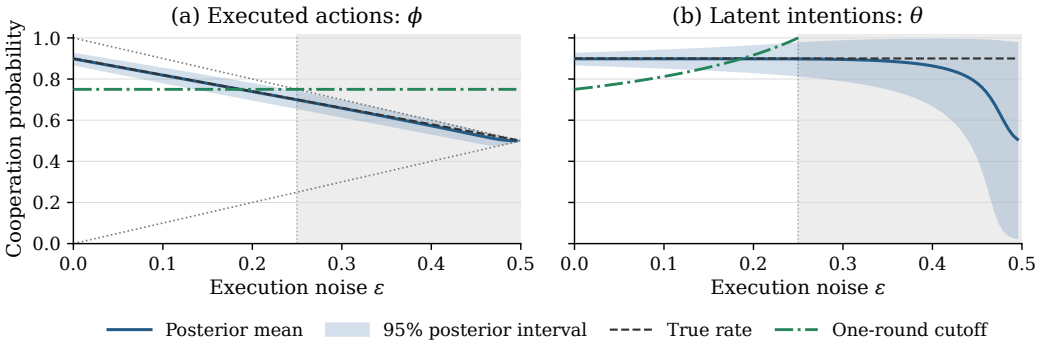


Figure 2: **Matched inference in observation and intent coordinates.** Stag Hunt $(R, T, P, S) = (4, 3, 3, 0)$; $\theta_{\text{true}} = 0.9$, $t = 500$. Both panels use the same representative count $k = t\phi_{\text{true}}$ at each ϵ and a uniform prior on θ (hence on $\phi \in [\epsilon, 1 - \epsilon]$). Bands are exact 95% equal-tailed posterior intervals. Green cutoffs describe the same one-round decision; dotted bounds mark attainable rates. Grey regions admit no strict cooperative one-round best response.

4.1 Bayesian model-based learning and planning

To this end, we study Bayesian model-based learners that select actions under their current beliefs and planning objectives. Game payoffs and execution noise are assumed known, while opponent dynamics are learned online. We contrast two architectural choices for this learning:

The **intent-conditioned learner** uses a partially observable Markov decision process (POMDP) over joint intended actions. It isolates the known execution channel using the BSC of §3.3 as its observation kernel Z, and receives payoffs on executed outcomes as reward. The opponent's reactive policy enters as an unknown transition kernel T. By contrast, the **observation-conditioned learner** dispenses with latent intent entirely, instead learning transitions directly between executed outcomes (Appendix A).

Let b_t denote state and model beliefs after the available history. The intent-conditioned learner approximates Bayesian learning through Dirichlet updates weighted by inferred state transitions; the observation-conditioned learner assigns evidence to observed transitions directly. Consequently, intent inference fundamentally alters which conditional predictions each observation updates.

4.2 Active inference and Bayes-adaptive planning

Given these models, the agents must select actions that balance immediate payoffs with information gathering. Active inference (AIF) achieves this by combining expected payoffs on executed outcomes with an explicit information value about states and transition parameters (Friston et al., 2017). Similarly, Bayes-adaptive planning values information through its consequences for future reward, where anticipated belief updates can improve subsequent action choices (Pineau et al., 2008).

Our primary comparison between AIF-POMDP and AIF-MDP fixes the payoff objective, horizon, and learning settings, isolating the effect of changing the state representation and channel placement. We also introduce an approximate Bayes-adaptive POMDP (BAPOMDP) baseline to extend this comparison to a different treatment of information and planning. Experiment-specific settings are detailed in Appendix E.

4.3 Decision boundaries and learning feedback

The critical link between these planning objectives and the behavioural trajectories we observe lies in the agents' decision boundaries. For a planner with horizon H and action scores S_H (where larger scores are preferred), the cooperation margin is defined as:

$$m_{t,H} = S_H(C \mid b_t, \epsilon) - S_H(D \mid b_t, \epsilon). \tag{5}$$

Positive values favour cooperation, negative values favour defection, and zero defines the decision boundary. Because these scores incorporate both future rewards and information value, a modest change in predictions near this boundary can reverse the agent's choice.

When noise raises the one-round threshold $b^\star_\epsilon$ (as in the Stag Hunt), the learner must support a stronger prediction of partner cooperation using observations that inherently reveal less about intent (§3.5). Appendix C quantifies the resulting decision-error probability for independent observations in a fixed context.

Crucially, what happens after a boundary crossing depends on how actions dictate subsequent learning. When an agent chooses defection, it experiences fewer cooperative situations. If its model learns separate predictions for different contexts, evidence collected while defecting does little to correct its prediction of what would happen following mutual cooperation. A pessimistic prediction can thereby persist simply because the agent rarely takes the actions required to test it.

Intent inference can reinforce this feedback loop: because the learned model dictates the evidentiary weight of each observation, a cooperative transition prediction can remain poorly tested even as total experience grows. We examine how these learning histories and action choices relate to cooperation collapse in the following experiments.

5 Empirical example: Stag Hunt against an unconditional cooperator

We examine the interaction between inference, action, and learning in a Stag Hunt $(R, T, P, S) = (4, 3, 3, 0)$ against a stationary unconditional cooperator (*AllC*). In this game, the noiseless cooperation threshold is $b^\star = 3/4$, while the noise-corrected threshold $b^\star_\epsilon = \frac{3-4\epsilon}{4(1-2\epsilon)}$ rises to 1 at $\epsilon = 1/4$. Because *AllC* always intends to cooperate, this setup provides a controlled environment to test whether the learner loses cooperation before reaching the theoretical noise limit for a known partner. We compare both representations across 300 seeds per noise level (5,000 rounds, planning horizon $H = 2$, model updates every 5 rounds), evaluating the cooperation rate over the final 1,250 rounds (Appendix F).

Intent-conditioned learning: a sharp loss of cooperation. Figure 3a shows a sharp change in the distribution of learned behaviour. Mean late cooperation falls from 0.927 at $\epsilon = 0.15$ to 0.201 at $\epsilon = 0.20$, while the median falls from 1 to 0.010. At $\epsilon = 0.20$, 203 of 300 seeds have cooperation rates at most 0.1, and 30 have rates at least 0.9, alongside intermediate outcomes. This change occurs while cooperation remains a strict one-round best response to a known AllC partner.

Observation-conditioned (MDP) agent: volatile noise-dependent clustering. The observation-conditioned learner spreads into several intermediate clusters (Figure 3b), with mean cooperation falling from 0.887 to 0.695 over the same noise interval. We hypothesize that this intermediate behavior occurs because agents learn specific conditional response rules, cooperating after certain executed outcomes and defecting after others.

To test this, we evaluate whether each run can be explained by a fixed policy over the four preceding outcomes (CC, CD, DC, DD). Using the first 625 rounds of the evaluation window, we assign each outcome the intended action that most frequently followed it. When we hold this extracted rule fixed, it predicts the learner's actual actions in the remaining 625 rounds with over 99.4% mean accuracy at $\epsilon \in \{0.15, 0.20, 0.25\}$.

Crucially, each extracted rule, combined with execution noise and *AllC*'s behavior, defines a transition matrix over the observed outcomes. The theoretical steady-state cooperation rates derived from these transitions match the empirically measured rates with a mean absolute error of just 0.5–1.6 percentage points. This confirms that the distinct clusters correspond to different learned response patterns: each pattern yields a characteristic cooperation rate, and noise shifts that rate by altering how often the outcomes that elicit cooperation occur (Appendix F).

Current intent attribution and conditional learning. The intent-conditioned learner maintains two distinct beliefs relevant to cooperation: a state belief about the partner's current intent, and a learned transition model predicting how joint intentions will evolve. Consequently, an agent might

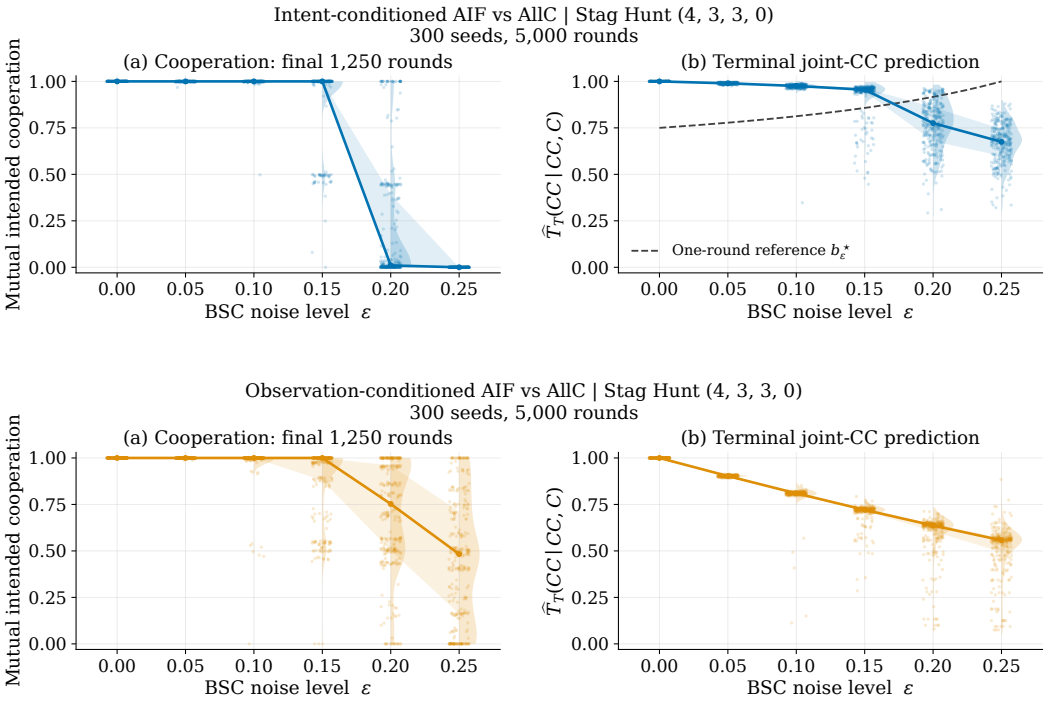


Figure 3: **Architectural comparison on Stag Hunt vs. AllC** (300 seeds per noise level, 5,000 rounds). Left: mutual intended cooperation over the final 1,250 rounds. Right: terminal learned joint-CC transition probability $\widehat{T}_T(CC \mid CC, C)$, from the normalised Dirichlet column. State labels denote intentions for the POMDP and executed outcomes for the MDP. Points denote seeds; lines and shading denote across-seed medians and interquartile ranges. The dashed line in the intent panel is the one-round reference $b_\epsilon^\star$, not the full AIF decision boundary.

correctly recognize that its partner is currently cooperating, yet still pessimistically predict that mutual cooperation is unsustainable.

We observe this disconnect by retrospectively analyzing independent trajectories at $\epsilon = 0.20$, comparing runs that ultimately collapse into defection (cooperating in $\leq 10\%$ of the final 1,250 rounds) against those that sustain it ($\geq 90\%$). At round 1,000, both groups hold nearly identical state beliefs, estimating roughly a 0.91 probability that the partner currently intends C. However, their transition models diverge sharply. Assuming both players previously intended C and the learner chooses C again, the high-cooperation group predicts a 0.923 probability that mutual cooperation will continue. In contrast, the low-cooperation group predicts only a 0.701 probability. For reference, the one-round threshold is $b_\epsilon^\star \approx 0.917$: the high-cooperation group lies just above it and the low-cooperation group well below. The full $H = 2$ AIF decision also includes continuation and information values.

Why do these predictions diverge? While both groups have accumulated the same total volume of transition evidence by round 1,000, the low-cooperation group has gathered far less evidence specific to sustained mutual cooperation. This is consistent with the feedback loop described in §4.3: choosing D reduces the opportunities to test the cooperative transitions used in planning (Appendix F.3).

In this configuration, parameter-information value enables cooperation to begin: removing it from initialisation leaves all 128 tested runs at each of $\epsilon = 0.15, 0.175, 0.20$ defecting throughout (Appendix F.7).

Cooperation–defection cycles near the boundary. The intermediate cluster in Figure 3a consists of learners that oscillate between cooperation and defection. At $\epsilon = 0.15$, the 36 (out of 300) runs that land in the $[1/3, 2/3]$ final-quarter cooperation range switch their intended action on 95.9% of rounds.

This alternating of C and D explains why these runs average out to roughly one-half cooperation, revealing that the population's sharp decline in cooperation actually masks highly dynamic, oscillatory behaviour within individual runs.

Planner diagnostics reveal the mechanics of this alternation. In a separate sample of intermediate runs at $\epsilon = 0.15$, the agent's action preference (cooperation margin) reliably flips based on its previous move: it leans negative (favouring D) after choosing C, and positive (favouring C) after choosing D. The agent's belief that the partner intends to cooperate remains steady at roughly 0.95 throughout. We hypothesize that the learner continuously crosses its decision boundary due to shifts in its inferred joint state. After choosing C, belief mass concentrates on the CC state, where the learner's transition model temporarily favours defection. After choosing D, the mass shifts to DC, making cooperation attractive again and so on (Appendix F.4).

This internal oscillation mirrors the classic retaliatory cycles seen in noisy Iterated Prisoner's Dilemmas, where accidental defections between Tit-for-Tat players trigger endless reciprocal retaliation (Au & Nau, 2006). However, because *AllC* never actually retaliates, the cycle here is entirely self-contained. The diagnostics suggest that changing joint-state beliefs and learned action values sustain this cycle even while the agent correctly perceives its partner's cooperative intent.

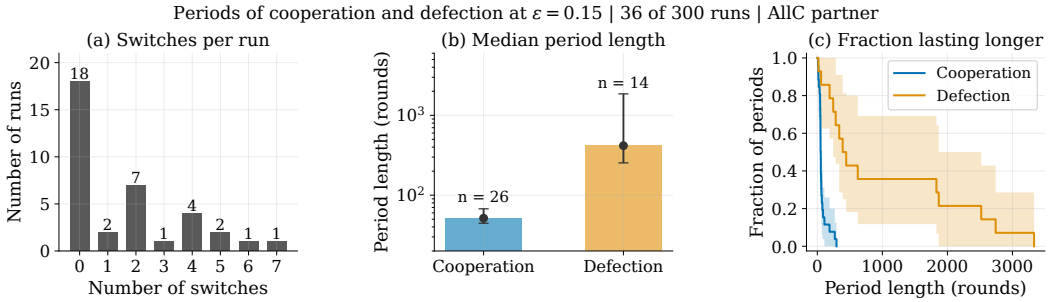


Figure 4: Periods of cooperation and defection at $\epsilon = 0.15$ (intent-conditioned AIF vs. *AllC*). The 36 selected runs have intended cooperation in $[1/3, 2/3]$ during the final 25% of the episode. A period is labelled as defection when the 50-round moving defection rate reaches 0.55, and as cooperation when it falls below 0.45; between these thresholds, the previous label is retained. (a) Switches per run over 5,000 rounds. (b) Median period lengths and interquartile ranges: 51.5 rounds for cooperation ($n = 26$) and 417 for defection ($n = 14$). (c) Fraction of completed periods lasting longer than the length on the horizontal axis, with shaded pointwise 95% seed-bootstrap intervals. Panels (b,c) exclude each run's first and last periods. Details and threshold sensitivity are in Appendix F.

6 Cross-game analysis

The cross-game experiments show that the sharp cooperation loss of §5 extends to reactive and co-learning partners, and that its behavioural consequences depend strongly on the game's payoffs. We compare Stag Hunt, Iterated Prisoner's Dilemma (IPD), Chicken, and Pure Coordination against Tit-for-Tat (TFT) and in self-play. Each architecture uses a fixed configuration across games and partners over $\epsilon \in \{0, 0.05, \ldots, 0.25\}$. Rates below describe intended actions over the final 1,250 rounds of 5,000-round episodes, averaged over 100 seeds, unless stated otherwise. The settings and per-game figures are in Appendices E and G.

Stag Hunt and IPD: sharp collapse extends beyond AllC. The architectural contrast becomes pronounced in self-play. Increasing noise from $\epsilon = 0.10$ to 0.15 reduces AIF-POMDP mutual cooperation from 0.813 to 0.110 in Stag Hunt and from 0.687 to 0.168 in IPD. At $\epsilon = 0.15$, AIF-MDP retains cooperation of 0.757 in both games. Thus, under the same AIF planning procedure and settings, the intent-conditioned representation exhibits a much sharper loss of cooperation. The seed distributions also retain the contrast between polarised outcomes and intermediate clusters seen against AllC (Appendix G).

Partner responses change robustness to noise. The Bayes-adaptive baseline also sustains substantially less cooperation in IPD self-play than against TFT: at $\epsilon = 0.15$, its mean mutual cooperation is 0.212 in self-play and 0.734 against TFT. TFT follows a fixed response rule, whereas both agents adapt in self-play. In IPD, defection dominates the stage game, so cooperation under a reward-based planner draws its value from expected future responses. These results show how strongly the cooperation sustained at a given noise level depends on the partner.

Chicken: lower cooperation accompanies asymmetric play. With horizon $H = 5$, AIF-POMDP maintains full intended mutual cooperation throughout the Chicken self-play sweep, including at noise levels where its Stag Hunt and IPD cooperation largely disappears. BAPOMDP reaches a different outcome: at $\epsilon = 0.25$, mutual cooperation accounts for 0.381 of rounds, asymmetric intentions for 0.600, and mutual defection for only 0.019. Its lower cooperation therefore reflects predominantly actions that are unilaterally cooperative. With Chicken payoffs $(R, T, P, S) = (3, 4, -1, 1)$, these asymmetric profiles avoid the worst outcome, DD. The one-round boundary also reverses direction relative to Stag Hunt: a lower predicted probability of partner cooperation favours cooperating to avoid mutual defection (Appendix D.2).

Pure Coordination: intended agreement remains high. Both intent-conditioned planners retain high agreement in self-play at $\epsilon = 0.25$: 0.990 for BAPOMDP and 0.913 for AIF-POMDP. Their realised agreement is much lower, approximately 0.625 and 0.605, respectively. Agents therefore continue to align their intended choices even when executed outcomes appear substantially less coordinated. Here $(R, T, P, S) = (1, 0, 1, 0)$ gives equal value to CC and DD, and the one-round boundary remains at $b^\star_\epsilon = 1/2$. Execution errors reduce even perfectly matched intentions to expected agreement $(1 - \epsilon)^2 + \epsilon^2 = 0.625$ at this noise level. Separating intended from realised agreement reveals the coordination retained behind the performance decline (Appendix D.3).

7 Discussion

Intent inference offers a way to distinguish execution errors from deliberate defection, but its benefit for cooperation depends on learning reliable predictions about the consequences of the agent's own actions. In Stag Hunt, intent-conditioned learners sharply abandon cooperation even against unconditional cooperators, despite maintaining high confidence in their partner's cooperative intent. We hypothesise that this collapse involves a feedback loop: pessimistic predictions discourage cooperation, which reduces the evidence available to revise them. Near a decision boundary, this feedback could amplify early learning fluctuations into sharply divergent behaviour.

Crucially, these consequences depend on the interaction between state representation, planning objective, and game payoffs. Intermediate cooperation rates can mask entirely distinct underlying mechanisms, ranging from outcome-conditioned response rules to rapid action alternation. Furthermore, cooperation loss manifests differently across games—shifting to asymmetric play in Chicken, while preserving intended agreement in Pure Coordination. Because population-level averages conflate these distinct patterns and obscure the impact of execution errors, evaluations in noisy settings must report per-run trajectories alongside both intended and executed joint actions.

While our analysis of this feedback loop relies on symmetric 2×2 games, known execution noise, and short planning horizons, the underlying vulnerabilities highlight a broad barrier to multi-agent coordination. Future work must investigate how deeper planning and richer opponent models affect evidence collection in more complex environments. The central design challenge is to enable agents to safely test and revise pessimistic predictions about cooperation, even when their current model strongly favours defection.

GENERATIVE AI DISCLOSURE

In this work, we used generative AI tools to assist with code, analysis of experiment outputs, preparation of plots, manuscript drafting and editing, and intermediary reviews and structural critiques. The figures report experiment outputs and analytical calculations.

We have reviewed all AI-assisted work. Specifically, LLM-generated code was verified and tested for correctness within our experimental pipeline, and all AI-drafted text, critiques, and suggestions were manually reviewed, edited, and evaluated for scientific accuracy by the authors. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Tsz-Chiu Au and Dana Nau. Accident or intention: That is the question (in the Noisy Iterated Prisoner's Dilemma). In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '06, pp. 561–568, New York, NY, USA, May 2006. Association for Computing Machinery. ISBN 978-1-59593-303-4. doi: 10.1145/1160633.1160736.

Robert Axelrod and William D. Hamilton. The Evolution of Cooperation. *Science*, 211(4489): 1390–1396, March 1981. doi: 10.1126/science.7466396.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, December 2009. ISSN 1873-7838. doi: 10.1016/j.cognition.2009.07.005.

Maarten C. Boerlijst, Martin A. Nowak, and Karl Sigmund. The Logic of Contrition. *Journal of Theoretical Biology*, 185(3):281–293, April 1997. ISSN 0022-5193. doi: 10.1006/jtbi.1996.0326.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open Problems in Cooperative AI, December 2020.

Daphne Demekas, Conor Heins, and Brennan Klein. An analytical model of active inference in the Iterated Prisoner's Dilemma, August 2023.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with Opponent-Learning Awareness, September 2018.

Jakob N. Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning, September 2019.

Karl J. Friston, Marco Lin, Christopher D. Frith, Giovanni Pezzulo, J. Allan Hobson, and Sasha Ondobaka. Active Inference, Curiosity and Insight. *Neural Computation*, 29(10):2633–2683, October 2017. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_00999.

Drew Fudenberg, David Levine, and Eric Maskin. The Folk Theorem with Imperfect Public Information. *Econometrica*, 62(5):997–1039, 1994. ISSN 0012-9682. doi: 10.2307/2951505.

Nikoleta E. Glynatsi, Vincent Knight, and Marc Harper. Properties of winning Iterated Prisoner's Dilemma strategies. *PLOS Computational Biology*, 20(12):e1012644, December 2024. ISSN 1553-734X. doi: 10.1371/journal.pcbi.1012644.

P. J. Gmytrasiewicz and P. Doshi. A Framework for Sequential Planning in Multi-Agent Settings. pp. 49–79, 2005. ISSN 10769757. doi: 10.1613/jair.1579.

John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA, USA, June 1988. ISBN 978-0-262-58238-4.

Christian Hilbe, Luis A. Martinez-Vaquero, Krishnendu Chatterjee, and Martin A. Nowak. Memory-n strategies of direct reciprocity. *Proceedings of the National Academy of Sciences*, 114(18): 4715–4720, May 2017. doi: 10.1073/pnas.1621239114.

Adrian Hutter. Balancing Cooperativeness and Adaptiveness in the (Noisy) Iterated Prisoner's Dilemma, March 2023.

Per Molander. The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution*, 29(4):611–618, 1985. ISSN 1552-8766. doi: 10.1177/0022002785029004004.

Martin Nowak and Karl Sigmund. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, 364(6432):56–58, July 1993. ISSN 1476-4687. doi: 10.1038/364056a0.

Joelle Pineau, S. Ross, and B. Chaib-draa. Bayes-Adaptive POMDPs: A New Perspective on the Explore-Exploit Tradeoff in Partially Observable Domains. In *International Symposium on Artificial Intelligence and Mathematics*, 2008.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind, March 2018.

Jaime Ruiz-Serra, Patrick Sweeney, and Michael S. Harré. Factorised Active Inference for Strategic Multi-Agent Interactions, November 2024.

Brian Skyrms. The Stag Hunt. *Proceedings and Addresses of the American Philosophical Association*, 75(2):31–41, 2001. doi: 10.2307/3218711.

Jianzhong Wu and Robert Axelrod. How to Cope with Noise in the Iterated Prisoner's Dilemma. *Journal of Conflict Resolution*, 39(1):183–189, March 1995. ISSN 0022-0027. doi: 10.1177/0022002795039001008.

A State representations and Bayesian learning

The two representations in §4.1 differ in how they use an observed outcome to learn about future play. Both learners know the game payoffs and execution-noise rate. The intent-conditioned learner assigns evidence to inferred transitions between joint intentions, while the observation-conditioned learner updates transitions between executed outcomes directly.

A.1 Intent-conditioned model

The learner represents a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, r, b_0)$. At decision time t, $s_t \in \{CC, CD, DC, DD\}$ describes the preceding joint intentions, with the learner's action listed first; an additional state s_0 marks initialization. The chosen action is $a_t \in \{C, D\}$ and o_t records the preceding executed outcome, with the same initial placeholder. Action a_t leads to s_{t+1} and observation o_{t+1}.

The learner maintains a separate prediction of the partner's next intention for each preceding state and chosen action. Because the learner knows its own intended action, the possible successor states satisfy

$$T(s' \mid s, a) = \mathbf{1}\{s'_i = a\}\, p_T(s'_j \mid s, a). \tag{6}$$

Here p_T predicts partner intent conditional on the represented state and chosen action. This model provides a finite-state approximation to the partner's responses. TFT responds to executed actions, while a self-play partner also changes its beliefs and policy; the learner uses its transition model to predict the resulting behaviour.

The fixed observation kernel applies independent execution flips:

$$Z(o \mid s) = \prod_{k \in \{i,j\}} \left[(1-\epsilon)\mathbf{1}\{o_k = s_k\} + \epsilon \mathbf{1}\{o_k \neq s_k\}\right]. \tag{7}$$

The initial placeholder is observed without noise. Rewards satisfy $r(CC), r(CD), r(DC), r(DD) = (R, S, T, P)$ and $r(s_0) = 0$. Thus the expected reward after action a is computed on executed outcomes,

$$\bar{r}(q, T, a) = \sum_{s, s', o} q(s) T(s' \mid s, a) Z(o \mid s') r(o). \tag{8}$$

A.2 Filtering and transition learning

The learner approximates its joint state and model belief by $q_t(s)\mu_t(T)$, with independent Dirichlet factors $\mu_t(T) = \prod_{s,a} \mathrm{Dir}(T(\cdot \mid s,a); N_{s,a})$ on the supported transitions. Filtering uses the current posterior-mean model $\bar{T}_t$:

$$q_{t+1}(s') \propto Z(o_{t+1} \mid s') \sum_s \bar{T}_t(s' \mid s, a_t) q_t(s), \qquad \bar{T}_t(s' \mid s, a) = \frac{N_{s,a}(s')}{\sum_v N_{s,a}(v)}. \tag{9}$$

Each observation contributes fractional transition counts $\Delta N_{s,a}(s') = \eta\, \mathbf{1}\{a_t = a\}\xi_t(s,s')$, where ξ_t approximates the inferred joint state transition and η is the learning rate. AIF accumulates these updates over the configured learning interval; the Bayes-adaptive approximation uses the product of consecutive filtered marginals (Appendix B.2). This approximate Bayesian update connects inference to learning: the current model helps determine which transition is inferred, and that inference determines which conditional prediction receives evidence.

A.3 Observation-conditioned model

The observation-conditioned learner takes executed outcomes as states, using $\widetilde{\mathcal{M}} = (\mathcal{O}, \mathcal{A}, \widetilde{T}, r, \widetilde{b}_0)$ with an identity observation kernel. Each observed transition contributes directly to the corresponding transition count:

$$\Delta \widetilde{N}_{o,a}(o') = \eta\, \mathbf{1}\{o_t = o,\ a_t = a,\ o_{t+1} = o'\}. \tag{10}$$

The learner knows its current state and uses its learned model to predict what follows each action. Its initial transition prior averages intended-action predictions through the known BSC. The two representations therefore incorporate knowledge of execution noise in different places: the observation-conditioned learner uses it in the transition prior, and the intent-conditioned learner retains it as a fixed observation likelihood.

When a stationary latent Markov kernel T describes the interaction, a visited observation–action pair has conditional next-outcome distribution

$$\Pr_\pi(o' \mid o, a) = \sum_{s,s'} Z(o' \mid s') T(s' \mid s, a) \rho_\pi(s \mid o, a), \tag{11}$$

where ρ_π describes the possible latent states conditional on the observed outcome and chosen action. With sufficient visits and suitable stationary sampling, direct transition counts estimate this next-outcome distribution. Whether the latest outcome contains enough information for prediction depends on the partner. Against AllC, the next executed outcome depends only on the chosen action and independent execution flips. For other partners, earlier history can provide additional information about the latent state. During adaptive play, the response distributions can also change as the agents change their policies.

B Planning implementations

Both planners use their current beliefs to evaluate the consequences of an action, but they value information differently. AIF explicitly rewards expected information about states and transition parameters. The approximate Bayes-adaptive planner anticipates how learning can improve future reward. This appendix specifies how each objective produces the action scores used in equation 5, given the models in Appendix A.

B.1 Active inference

AIF evaluates action sequences of length H under the current posterior-mean transition model. For a sequence π, predicted state and observation distributions satisfy $q^\pi_{k+1} = \bar{T}^\mathsf{T}_{\pi_k} q^\pi_k$ and $q^\pi_{k+1}(o) = \sum_s Z(o \mid s) q^\pi_{k+1}(s)$, starting from the filtered state belief. The negative expected free energy is the policy score

$$F_H(\pi) = \sum_{k=1}^{H} \left[\mathbb{E}_{q^\pi_k(o)} r(o) + I^\pi_{s,k} + I^\pi_{T,k} \right]. \tag{12}$$

The reward term uses the game payoffs as observation preferences. Equivalently, a preferred-outcome distribution proportional to $\exp r(o)$ gives the same policy comparison through its log probabilities, up to a common additive constant. The state-information term is

$$I^{\pi}_{s,k} = \mathcal{H}[q^{\pi}_k(o)] - \sum_s q^{\pi}_k(s)\mathcal{H}[Z(\cdot \mid s)], \tag{13}$$

where $\mathcal{H}$ denotes entropy. It values information that a future observation would provide about the corresponding hidden state (Friston et al., 2017).

The parameter-information term rewards learning about transitions that remain uncertain. We approximate this value using weights determined by the accumulated transition evidence:

$$I^{\pi}_{T,k} = \sum_{s,s'} q^{\pi}_{k-1}(s) q^{\pi}_k(s')\, w_{s,\pi_{k-1}}(s'), \qquad w_{s,a}(s') = \frac{1}{N_{s,a}(s')} - \frac{1}{\sum_v N_{s,a}(v)}. \tag{14}$$

Weights apply only to supported entries with positive concentration. Counts and transition means remain fixed while each sequence is evaluated, so this term approximates the value of future learning using current uncertainty. Online transition updates follow Appendix A.

With a uniform prior over candidate sequences, $q_t(\pi) \propto \exp\{\gamma F_H(\pi)\}$, where $\gamma > 0$ is policy precision. To choose an action, the planner sums the probabilities of all sequences that begin with that action and selects the larger sum. The corresponding action score is

$$S_H(a \mid b_t, \epsilon) = \gamma^{-1} \log \sum_{\pi:\pi_0=a} q_t(\pi). \tag{15}$$

Ties select C. Reward and information terms add at the sequence level; summing over sequences and taking the logarithm means that their separate action margins do not generally add. AIF-MDP uses the same calculation with observed states and an identity likelihood. As transition beliefs concentrate, the parameter-information bonus diminishes. Future rewards and, for the intent-conditioned learner, information about hidden states can still influence the action choice.

B.2 Approximate Bayes-adaptive planning

The Bayes-adaptive baseline considers how possible observations would change the learner's model and subsequent decisions (Pineau et al., 2008). Each planning node contains a filtered state distribution q and a table of accumulated transition evidence N. Using the local Dirichlet mean $\bar{T}_N$, the planner considers each possible next observation and updates both beliefs and counts along that branch. For chosen action a and hypothetical observation o, the fractional count update is

$$N^{a,o}_{s,a}(s') = N_{s,a}(s') + q(s) q^{a,o}(s'), \tag{16}$$

with other action columns unchanged. Online learning uses the same product of consecutive state beliefs. A single table at each node summarises the possible count histories, giving an approximation to their joint posterior with the hidden state.

At depth $d \geq 1$, the recursion is

$$V_d(q, N) = \max_a \left\{ \bar{r}(q, \bar{T}_N, a) + \delta \sum_o \Pr(o \mid q, N, a) V_{d-1}(q^{a,o}, N^{a,o}) \right\}, \tag{17}$$

where δ discounts reward and equation 8 defines $\bar{r}$. At a leaf, a configured number of value-iteration sweeps under the local mean model gives $Q_N(s, a)$, and $V_0(q, N) = \max_a \sum_s q(s) Q_N(s, a)$. Hypothetical count updates can therefore change the continuation value. The planner adds no explicit information bonus.

The action with the largest value at the root is selected, with ties assigned to C. The learner reuses these action values between replanning steps. The cross-game configuration uses depth one, one leaf sweep, and replanning every five rounds (Appendix E.3). The reported baseline therefore combines shallow planning over possible learning updates with a continuation value from the updated mean model.

C STAG HUNT DECISION-MARGIN REFERENCE

We quantify how sampling uncertainty can lead a learner to defect when the true partner cooperation rate favours cooperation. For a one-round decision based on independent observations in a fixed context, the error probability depends on the distance to the decision boundary and the uncertainty in the estimated intent rate. This calculation provides a reference for interpreting the learning dynamics in §5.

C.1 ASSUMPTIONS AND SAMPLING UNCERTAINTY

Fix a partner intent rate θ_{true} and a known noise rate $\epsilon \in [0, 1/2)$. Let $X_1, \ldots, X_t$ be independent indicators of observed partner cooperation in that context, with mean $\phi_{\text{true}} = \epsilon + (1 - 2\epsilon)\theta_{\text{true}}$. The channel-inversion estimator is $\hat{\theta}_t = (\bar{X}_t - \epsilon)/(1 - 2\epsilon)$. It is unbiased, with exact sampling variance

$$\sigma_\epsilon^2(t) = \frac{\phi_{\text{true}}(1 - \phi_{\text{true}})}{t(1 - 2\epsilon)^2}. \tag{18}$$

For $0 < \phi_{\text{true}} < 1$ and sufficiently large t, the estimator's distribution across independent datasets is approximately $\mathcal{N}(\theta_{\text{true}}, \sigma_\epsilon^2(t))$. Because channel inversion is unconstrained, an estimate can lie outside $[0, 1]$. The variance describes how estimates differ across repeated samples. A Bayesian credible interval, such as those in Figure 2, instead describes uncertainty conditional on one observed sample and its prior.

The reference rule cooperates iff $\hat{\theta}_t \geq b_\epsilon^\star$, equivalently $\bar{X}_t \geq b^\star$, assigning ties to cooperation. We consider an attainable interior Stag Hunt cutoff and a sample size t fixed in advance. Here each sample supplies independent evidence about the same context. For an online learner, the relevant evidence depends on which contexts its actions visit and how inference assigns observations to them.

C.2 A CONDITIONAL DECISION-ERROR CALCULATION

Proposition 1 (Scalar decision-error reference)**.** *Under the preceding sampling model and decision rule, the probability of choosing defection at the evaluation point is*

$$p_D(t, \epsilon) = \sum_{k=0}^{\lceil tb^\star \rceil - 1} \binom{t}{k} \phi_{\text{true}}^k (1 - \phi_{\text{true}})^{t-k}. \tag{19}$$

When $\Delta(\epsilon) = \theta_{\text{true}} - b_\epsilon^\star > 0$, this is a decision error relative to the one-round action preferred under the true rate. Its normal approximation is

$$p_D(t, \epsilon) \approx \Phi\left(-\frac{\Delta(\epsilon)}{\sigma_\epsilon(t)} \right). \tag{20}$$

Here Φ is the standard normal cumulative distribution function.

Proof. Defection is selected iff $K = \sum_i X_i \leq \lceil tb^\star \rceil - 1$. Since $K \sim \text{Binomial}(t, \phi_{\text{true}})$, summing these probabilities gives the exact expression. Standardising the estimator in equation 18 gives the normal approximation. The probability concerns a single evaluation point across repeated datasets. □

For $0 < \alpha < 1$, comparison with a two-sided $1 - \alpha$ normal interval uses $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. The margin equation

$$\Delta(\varepsilon^\star) = z_{\alpha/2}\, \sigma_{\varepsilon^\star}(t) \tag{21}$$

marks an approximate *one-sided* decision-error probability $\alpha/2$. We use $\varepsilon^\star$ for this reference noise level when a root exists in the cooperative region. It locates where sampling uncertainty makes an incorrect one-round choice more likely. How that choice affects subsequent cooperation depends on the resulting learning and planning. Because channel inversion transforms both the estimate and its cutoff, the matched one-round rule has the same error probability in intent and observation coordinates.

C.3 Symmetric Stag Hunt $(R, T, P, S) = (3, 2, 1, 0)$

Here $R - T = P - S = 1$, so $b^\star_\epsilon = 1/2$. Against an unconditional cooperator, $\theta_{\text{true}} = 1$ and $\phi_{\text{true}} = 1 - \epsilon$. Equation equation 21 becomes

$$t(1 - 2\varepsilon^\star)^2 = 4z^2_{\alpha/2}\varepsilon^\star(1 - \varepsilon^\star), \qquad \varepsilon^\star = \frac{1}{2}\left(1 - \frac{z_{\alpha/2}}{\sqrt{t + z^2_{\alpha/2}}}\right). \tag{22}$$

There is a unique root in $(0, 1/2)$, approaching $1/2$ as the fixed number of independent observations grows. For $\alpha = 0.05$ ($z \approx 1.96$), it is 0.404 at $t = 100$ and 0.456 at $t = 500$.

C.4 Asymmetric experimental Stag Hunt $(R, T, P, S) = (4, 3, 3, 0)$

The threshold is $b^\star_\epsilon = (3 - 4\epsilon)/[4(1 - 2\epsilon)]$, rising from $3/4$ to 1 as ϵ reaches $1/4$. Against AllC the margin is $\Delta(\epsilon) = (1 - 4\epsilon)/[4(1 - 2\epsilon)]$. Thus

$$t(1 - 4\varepsilon^\star)^2 = 16z^2_{\alpha/2}\varepsilon^\star(1 - \varepsilon^\star). \tag{23}$$

There is a unique root in $(0, 1/4)$. For $\alpha = 0.05$ and $t = 500$, it is $\varepsilon^\star \approx 0.214$, above the observed cooperation-loss interval 0.15–0.20. The empirical learner therefore loses cooperation earlier than this fixed-sample calculation suggests. Its conditional predictions, allocation of evidence, and planning objective are the additional ingredients examined in §5. At $\epsilon = 1/4$, cooperation ceases to be a strict one-round best response even when AllC's policy is known.

C.5 Scope across games and learners

For Stag Hunt payoffs, substituting the threshold from equation 3 into the margin equation gives the same calculation wherever the cutoff is attainable and the cooperative margin is positive. Other games can reverse the direction of the preferred response or eliminate an interior cutoff (Appendix D). For a finite-horizon learner, the relevant boundary is the full action comparison in equation 5, which includes future rewards and information incentives. Using the sampling calculation to predict that learner's choices requires establishing how its estimates, evidence count, and planning objective relate to the one-round rule.

D Cross-game decision boundaries

The consequences of a changing cooperation prediction depend on the game. The payoff comparison in §3 determines the preferred immediate response; the action margin in equation 5 also includes future rewards and information. Here we derive the comparisons that help explain the cross-game results: future responses can give cooperation value in IPD, Chicken rewards cooperation against a likely defector, and Pure Coordination rewards agreement on either action.

D.1 Horizon dependence in the IPD

The IPD payoffs $(R, T, P, S) = (3, 5, 1, 0)$ satisfy $T > R > P > S$, so defection dominates the stage game. Under a reward-based planning objective, cooperation can become preferable through the partner responses expected over subsequent rounds. The relevant comparison therefore depends on the learned conditional response model and the continuation policy.

TFT responds to the learner's previous executed action and can return to cooperation after the learner cooperates again. Execution noise affects both immediate payoffs and these future responses. In self-play, the partner's changing beliefs also affect its actions. Evaluating cooperation therefore requires the planner's full action comparison in equation 5, including the continuation policies it considers.

D.2 Chicken

Chicken payoffs $(R, T, P, S) = (3, 4, -1, 1)$ satisfy $T > R > S > P$. The payoff advantage of cooperation decreases as the predicted probability of partner cooperation increases: cooperating

avoids mutual defection when the partner is likely to defect, while defection gains the temptation payoff when the partner is likely to cooperate. Substitution into equation 3 gives

$$b^\star_\epsilon = \frac{2 - 3\epsilon}{3(1 - 2\epsilon)}, \tag{24}$$

which rises from $2/3$ at $\epsilon = 0$ to $5/6$ at $\epsilon = 0.25$. The one-round comparison favours cooperation for $b < b^\star_\epsilon$ and defection for $b > b^\star_\epsilon$. Thus Chicken has an interior boundary over the tested range, with the opposite orientation to Stag Hunt.

The asymmetric pure equilibria (CD) and (DC), together with the interior mixed equilibrium, illustrate why lower mutual cooperation has a different meaning in Chicken. If the other player continues to defect, cooperation gives $S = 1$ rather than the mutual-defection payoff $P = -1$. Against TFT, the learner must also consider how its executed action changes the partner's next response. We therefore interpret the learned behaviour through joint-action frequencies: a reduction in mutual cooperation can accompany more asymmetric play while mutual defection remains rare.

D.3 Pure Coordination

Pure Coordination payoffs $(R, T, P, S) = (1, 0, 1, 0)$ satisfy $R = P > S = T$. The two matching profiles give the same reward. Since $R - T = P - S = 1$, the one-round cutoff is

$$b^\star_\epsilon = b^\star = 1/2, \qquad 0 \le \epsilon < 1/2. \tag{25}$$

Both players choosing C and both choosing D achieve the same intended agreement. During learning, agreement falls when their predictions lead to different choices or when they change actions without coordinating. Mutual cooperation alone therefore misses successful coordination on D.

Let q denote the probability of matching intended actions before independent execution flips. Executed actions match with probability $(1 - \epsilon)^2 + \epsilon^2$ when intentions match, and $2\epsilon(1 - \epsilon)$ when intentions differ. Hence

$$\Pr(\text{executed agreement}) = 2\epsilon(1 - \epsilon) + (1 - 2\epsilon)^2 q. \tag{26}$$

At $q = 1$, this reduces to $(1 - \epsilon)^2 + \epsilon^2$, which is 0.625 at $\epsilon = 0.25$. Thus even perfectly aligned intentions produce substantial executed disagreement under noise. Reporting both quantities shows how much coordination the agents retain and how much reward they lose through execution errors; in this game, the executed agreement rate is also the expected stage-game payoff.

E Experimental setup and reproducibility

E.1 Environment and evaluation protocols

The experiments compare intended behaviour under a shared execution-noise process. Flips occur independently across players and rounds, and executed actions determine the observations and payoffs. The agents know the environmental flip rate. The main sweeps use $\epsilon \in \{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$. We measure intended mutual cooperation over the final quarter of each episode, or agreement on either intended action in Pure Coordination.

Experiment	Rounds	Seeds per noise	Partner
IPD tuning	1,000	5	Self-play
IPD lead figure	3,000	200	Self-play
Stag Hunt–AllC benchmark	5,000	300	AllC
Cross-game comparison	5,000	100	TFT and self-play

Self-play uses two separately learning agents. AllC always intends C; TFT responds to the other player's previous executed action. Hyperparameter tuning uses IPD payoffs $(R, T, P, S) = (3, 5, 1, 0)$ and selects by mean mutual cooperation across seeds and noise levels. The cross-game comparison fixes one configuration per architecture across games and partners. The AllC benchmark has a separate AIF horizon and learning interval.

E.2 SHARED AIF SETTINGS

The AIF comparison holds the planner, learning rate, policy precision, and prior strength fixed while changing the state representation and placement of the execution channel. AIF-POMDP learns over joint intentions with a fixed noisy observation likelihood. AIF-MDP learns over directly observed outcomes and incorporates the channel in its transition prior (Appendices A and B).

Table 1: AIF tuning grid and evaluation settings, shared by both representations.

Parameter	Tuning values / fixed settings	IPD / cross-game	AllC
Planning horizon	$\{1, 3, 5\}$	5	2
Transition-update interval	$\{1, 5, 10, 50, 70, 100\}$	50	5
Prior concentration scale	Fixed at 1	1	1
Transition learning rate	Fixed at 1	1	1
Policy precision γ	Fixed at 1	1	1
Action selection	Deterministic	Deterministic	Deterministic

Transition learning is enabled throughout. Initial intent-transition means assign equal probability to the two successors compatible with the chosen action. To verify reproducibility, we repeated the AllC runs for seeds 0–3 at $\epsilon = 0.15, 0.20$ in both representations (16 trajectories), recovering identical actions, outcomes, and transition concentrations over 5,000 rounds. Recording beliefs and planner scores also preserved actions and updates in a four-seed, 500-round check per representation; the recorded policy-score components summed to the total score within 1.91×10^{-6}.

E.3 BAYES-ADAPTIVE SETTINGS

The approximate planner in Appendix B.2 uses soft count updates and a posterior-mean continuation value. Its discount factor is denoted δ, to distinguish it from AIF policy precision.

Table 2: Bayes-adaptive tuning grid and cross-game configuration.

Parameter	Tuning values	Cross-game value
Planning depth	$\{1, 2, 3\}$	1
Replanning interval	$\{1, 5\}$	5
Reward discount δ	$\{0.5, 0.9, 0.95, 0.99\}$	0.9
Prior concentration scale	$\{0.1, 1\}$	0.1
Prior bias	—	0.3
Leaf value-iteration steps	—	1
Count update	—	Soft

Prior bias sets the initial probabilities of CC after choosing C and DD after choosing D; their complementary probabilities go to CD and DC, respectively. It therefore specifies conditional response predictions rather than a single unconditional cooperation rate. The table distinguishes tuned parameters from those held fixed throughout the sweep.

E.4 Q-LEARNING SETTINGS

The Q-learning baseline in Appendix G learns from executed outcomes and received rewards. It uses an ϵ-greedy selector with multiplicative decay and optimistic initialization.

E.5 INDEPENDENT ALLC SAMPLES

The supplementary AIF analyses use the horizon-two AllC configuration above. We generate separate development and test samples containing 64 and 256 seeds per representation and noise level at $\epsilon = 0.10, 0.15, 0.175, 0.20, 0.225$, each run for 5,000 rounds. Generator seeds 2026092601 and 2026092602 define the two samples, with runs indexed by noise and replicate. Corresponding runs across representations share execution-noise draws. These independent samples test whether

Table 3: Q-learning tuning grid and selected configuration. Mean tuning cooperation is 0.65, averaged over five seeds and six noise levels.

Parameter	Tuning values	Selected value
Initial exploration ϵ	$\{0.05, 0.10\}$	0.05
Exploration decay	$\{1, 0.999, 0.99\}$	0.999
Minimum exploration	$\{0\}$	0
Learning rate	$\{0.01, 0.1, 0.5\}$	0.01
Reward discount	$\{0.5, 0.9, 0.995\}$	0.995
Initial Q-value	$\{50\}$	50

the conditional-learning patterns, alternation, and cooperation forecasts extend beyond the original trajectories (Appendix F).

The information ablation uses 128 paired seeds at each of $\epsilon = 0.15, 0.175, 0.20$. The cooperation probe uses a separate 256-seed sample at each of $\epsilon = 0.175, 0.20, 0.225$, generated with seed 2026092603. The selection criteria, controls, and uncertainty estimates accompany the results in Appendices F.7 and F.8.

F Supporting analyses of cooperation and learning

The following analyses connect the AllC cooperation distributions to the learners' response rules, beliefs, and opportunities to learn. We first examine the behaviour underlying the clusters, then test whether early beliefs predict later cooperation and whether additional evidence changes it. Each analysis identifies whether it uses the original 300-seed sample or the independent samples in Appendix E.5.

F.1 What the learned transition probability measures

For concentrations $N_{s,a}(s')$, the right panels of Figure 3 show the terminal ratio $N_{CC,C}(CC)/\sum_{s'} N_{CC,C}(s')$. This is the learner's prediction that CC will follow when the previous state was CC and it chooses C again. CC denotes joint intentions for AIF-POMDP and joint executed outcomes for AIF-MDP. In the latter case, this is the probability that both executed actions are C; it includes execution errors in the learner's own action. Points show individual runs and bands show variation across seeds. The displayed probability is one of the conditional predictions used in planning.

F.2 Outcome-conditioned response rules and intermediate clusters

To explain the observation-conditioned clusters, we test whether each run settles into a consistent response to the previous executed outcome during evaluation. For each seed, we fit a deterministic rule $f : \{CC, CD, DC, DD\} \to \{C, D\}$ using the most frequent intended action after each outcome in rounds 3,751–4,375. We then hold the rule fixed and predict actions in rounds 4,376–5,000. All four contexts occur in the fitting window for all 300 observation-conditioned seeds at each of $\epsilon = 0.15, 0.20, 0.25$.

Against AllC, execution noise determines how often the learner encounters each outcome. Following intended C and D, respectively, the outcome probabilities in the order CC, CD, DC, DD are

$$r_C = ((1-\epsilon)^2, (1-\epsilon)\epsilon, \epsilon(1-\epsilon), \epsilon^2), \qquad r_D = (\epsilon(1-\epsilon), \epsilon^2, (1-\epsilon)^2, (1-\epsilon)\epsilon).$$

Write $\alpha = \sum_s r_C(s)\mathbf{1}\{f(s) = C\}$ and $\beta = \sum_s r_D(s)\mathbf{1}\{f(s) = C\}$. These are the probabilities that the fixed rule chooses C after the learner previously chose C and D, respectively. Its stationary intended cooperation rate is $\beta/(1-\alpha+\beta)$ when the denominator is positive. The table compares the rule's predicted actions and stationary rate with the later observed actions and cooperation rate. Rate MAE is the mean absolute error across seeds.

ϵ	Mean action accuracy	Seeds with accuracy ≥ 0.99	Rate MAE
0.15	0.99834	296/300	0.00497
0.20	0.99554	282/300	0.01294
0.25	0.99421	282/300	0.01589

These comparisons include every seed. The fitted rules closely predict both individual actions and average cooperation in the subsequent window. This explains the clusters as different learned response patterns: each rule has a characteristic cooperation rate, which changes with noise as the frequencies of its triggering outcomes change. The stationary calculation describes these fixed rules; convergence of the continuing learning process would require a separate long-run analysis.

F.3 Current intent beliefs and learned cooperative contexts

A belief that the partner currently intends C answers a different question from a prediction that mutual cooperation will continue after choosing C. To examine this distinction, we group the independent test trajectories by final-quarter cooperation: low (≤ 0.10), high (≥ 0.90), or intermediate (Appendix E.5). At $\epsilon = 0.20$, there are 180 low, 23 high, and 53 intermediate trajectories. The table compares the two extreme groups. Current-belief and action summaries average rounds 951–1,000; transition-model summaries use the counts after round 1,000. The grouping is retrospective, using later behaviour to examine earlier learning histories.

Diagnostic at $\epsilon = 0.20$	Later low C	Later high C
Recent intended cooperation	0.085	0.566
Filtered current partner-cooperation probability	0.909	0.911
Concentration of (CC, C) transition column	36.9	361.0
Mean joint-CC probability in that column	0.701	0.923
Cumulative evidence following intended C	153.3	486.8
Total accumulated transition evidence	995	995

Both groups recognise the partner's cooperative intent, yet the group that later defects predicts a much lower chance of sustained mutual cooperation. Its lower concentration in the cooperative transition column means that this prediction has received less evidence, even though total transition evidence is the same. Figure 6(a,b) traces how these differences develop. The comparison supports the proposed link between behaviour and conditional learning; the intervention below tests whether supplying additional cooperative evidence changes subsequent behaviour.

F.4 Rapid alternation and conditional planner preferences

We examine intended actions directly to explain the intermediate cooperation cluster. At each noise level, we select runs whose cooperation over rounds 3,751–5,000 lies in $[1/3, 2/3]$. For each selected run, we compute the fraction of the 1,249 successive round pairs within this window on which the intended action changes. Averaging these fractions gives 0.9586 for the 36 selected runs at $\epsilon = 0.15$ and 0.8215 for the 46 selected runs at $\epsilon = 0.20$ in the 300-seed dataset of Figure 3. At $\epsilon = 0.15$, 20 of the 36 runs change action on more than 98% of successive pairs. Rapid alternation thus gives a direct behavioural explanation for cooperation rates near one-half.

To examine why the action preference changes, we analyse the independent test trajectories in Appendix E.5. This analysis is exploratory, motivated by the observed alternation. The same selection rule identifies 32 of 256 runs at $\epsilon = 0.15$ and 41 of 256 at $\epsilon = 0.20$. For every decision in the final 1,250 rounds, we record the filtered state belief and cooperation margin, defined by equation 5 with the AIF action score in equation 15. We group decisions by the learner's preceding intended action, then pool them to calculate the action frequencies, median margins, and mean beliefs below.

Diagnostic	$\epsilon = 0.15$	$\epsilon = 0.20$
Selected runs	$32/256$	$41/256$
Fraction of successive pairs changing action	0.9797	0.8153
Fraction choosing D after own C	0.9994	0.9173
Fraction choosing C after own D	0.9608	0.7336
Median cooperation margin after own C	-0.05846	-0.01127
Median cooperation margin after own D	$+0.02872$	$+0.00295$
Mean belief in partner C after own C	0.9574	0.9429
Mean belief in partner C after own D	0.9529	0.9195

The learner continues to attribute cooperative intent to its partner, while its joint-state belief shifts mainly between CC after choosing C and DC after choosing D. Its preferred action changes with that state: the median margin favours D after C and C after D. This supports the hypothesis that state-dependent predictions and information value repeatedly move the learner across its decision boundary. The result establishes alternation during the evaluation window; its eventual persistence remains open.

F.5 Switching between periods of predominant cooperation and defection

Figure 4 tracks which action predominates over longer windows. We smooth the intended defection indicator using a 50-round moving average, label the phase D-dominant when the average reaches 0.55, and return to C-dominant below 0.45. The first smoothed value determines the initial label. A switch is a change of label, and repeated switching means at least two changes. At $\epsilon = 0.15$, the 36 intermediate-cooperation seeds among the original 300 contain 18 with no switches, two with one, and 16 with at least two. Excluding the first and last phases, whose full durations may extend beyond the observation window, leaves 26 complete C-dominant and 14 complete D-dominant phases. These durations describe the selected completed phases and give more weight to seeds that switch more often.

Window	Enter / leave D	Any switch	At least two
25	$0.55/0.45$	292	269
50	$0.55/0.45$	215	144
100	$0.55/0.45$	131	47
50	$0.70/0.30$	75	43
50	$0.90/0.10$	15	1

The table varies the smoothing window and the thresholds across all 300 intent-conditioned seeds at $\epsilon = 0.15$. Modest changes in the dominant action are common, while repeated shifts between near-pure cooperation and defection are rare under these definitions. Rapid C–D alternation can keep a 50-round average near one-half without changing its phase label. These counts therefore describe a slower timescale than the round-by-round alternation in §F.4, and they include the full episode from early learning onward.

F.6 Predicting later cooperation from early learning

We next test whether early beliefs and planning preferences help predict which runs will eventually show low cooperation. The independent development and test samples in Appendix E.5 allow us to select the prediction procedure before evaluating it on new trajectories.

The outcome is intended cooperation at most 0.10 over rounds 3,751–5,000. We consider making the prediction at rounds $25, 50, 100, 250, 500$. A behaviour-based predictor uses the noise level and cooperation over the preceding 25 rounds. An extended predictor also uses the current action margin, its recent mean and standard deviation, the filtered probability of partner cooperation, and evidence about sustained mutual cooperation. The latter consists of log-one-plus concentration in the transition column for previous CC and chosen C, denoted (CC, C), and recent evidence assigned to that column. All features are available at the prediction round: scores use the pre-action model and filtered state, and concentrations include updates completed by that round. The action margin uses equation 15.

We assess predictions using the Brier score, the mean squared difference between predicted event probability and observed outcome; lower scores indicate better forecasts. Four-fold development cross-validation groups trajectories by seed index across noise levels. We select the earliest prediction round at which adding the belief and planning features reduces the intent-conditioned Brier score by at least 0.002, with a pre-specified fallback of round 50. The development differences, extended minus behaviour-based, are $-0.00137, -0.00637, -0.00757, -0.00759, -0.01024$ at the five candidate rounds. This selects round 50 for both architectures. Logistic models use development-only standardisation and minimise summed negative log likelihood plus half the squared coefficient norm, with an unpenalised intercept. A noise-only baseline estimates the event frequency at each noise level with Laplace smoothing. The prediction round, coefficients, and analysis procedure are fixed before generating the test trajectories.

Independent-test Brier score	Intent-conditioned	Observation-conditioned
Noise only	0.05479	0.03384
Noise and recent cooperation	0.04637	0.03163
Plus margin, belief, and evidence features	0.04783	0.03306

Late cooperation meets the low-cooperation criterion in 452 and 46 of the 1,280 test trajectories per architecture, respectively. Adding belief and planning features changes the Brier score by 0.00146 (95% bootstrap interval $[-0.00236, 0.00534]$) and 0.00143 ($[0.00016, 0.00286]$). Intervals use 2,000 paired resamples of test seed indices across noise levels, keeping the fitted models fixed. The additional features therefore provide no improvement on independent seeds, despite their apparent benefit during development.

The independent sample nevertheless reproduces the architectural contrast: intent-conditioned mean late cooperation is $0.996, 0.935, 0.733, 0.185, 0.001$ across the five noise levels, compared with $0.985, 0.904, 0.847, 0.700, 0.596$ for observation-conditioned learning (Figure 5). Among the 19 intent-conditioned and 266 observation-conditioned test trajectories with at least 0.80 cooperation in rounds 26–50, none later meets the low-cooperation criterion. The forecast therefore evaluates eventual learning outcomes; this sample contains no examples of collapse after cooperation is already established at the prediction round.

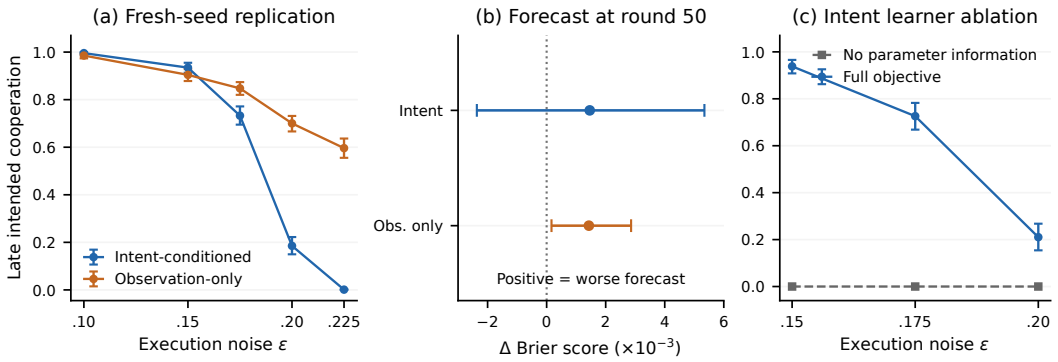


Figure 5: **Independent validation in the horizon-two AllC setting.** (a) Mean final-quarter intended cooperation, 256 independent test seeds per noise and architecture. (b) Test Brier change from adding planner-margin, belief, and evidence features to noise and recent cooperation at round 50; positive values are worse. (c) Parameter-information ablation from initialization, 128 paired intent-conditioned seeds per noise. Bars are 95% seed-bootstrap intervals, paired in (b). These samples are separate from Figure 3.

Earlier exploratory forecast. An earlier analysis of the original AllC sample predicted at round 1,000 whether cooperation would remain at most 0.20 in each of four 250-round blocks spanning rounds 4,001–5,000. Even and odd seed indices supplied 900 training and 900 test trajectories per representation. A logistic predictor used noise indicators and cooperation over rounds 501–1,000; a second added the cooperative-column mean and log concentration at round 1,000, with the

regularisation and training-only scaling described above. Adding these features changed test Brier score by 0.00246 (95% interval $[-0.00053, 0.00559]$) for intent-conditioned learning and -0.000029 ($[-0.00096, 0.00087]$) for observation-conditioned learning. Intervals use 1,000 paired resamples of test seed indices, keeping noise levels together. Here too, no trajectory with at least 0.80 recent cooperation met the later low-cooperation criterion. Both analyses thus concern predicting eventual low cooperation within the sampled learning histories.

F.7 Information value and opportunities to learn

To test how information value affects the collection of cooperative evidence, we remove AIF's parameter-information term from initialization at $\epsilon = 0.15, 0.175, 0.20$. The protocol is specified before inspecting intervention outcomes. State-information value, payoff preferences, priors, inference, learning, and action selection remain unchanged. The comparison uses 128 seeds per noise, sharing execution-noise draws with the first 128 independent-test baseline trajectories. It isolates one information term within AIF; Bayes-adaptive planning uses the different procedure described in Appendix B.2.

ϵ	Full-objective late C	Ablated late C	Paired difference, 95% interval
0.15	0.93822	0	-0.93822 $[-0.96577, -0.90831]$
0.175	0.72599	0	-0.72599 $[-0.78250, -0.66856]$
0.20	0.21036	0	-0.21036 $[-0.26774, -0.15418]$

Intervals use 5,000 paired seed-bootstrap resamples, with difference defined as ablated minus full. Every ablated trajectory defects throughout all 5,000 rounds. The full objective's first-500-round cooperation means are $0.662, 0.461, 0.270$. Mean accumulated evidence assigned to transitions after intended C is approximately $4{,}469$, $3{,}335$, and $1{,}045$ under the full objective and zero under ablation; total transition evidence is approximately $4{,}995$ in both arms. Thus action selection changes which transitions receive evidence, even when total evidence is unchanged.

Under these priors and planning settings, parameter-information value initiates cooperative exploration and supplies evidence about what follows cooperation. Since its removal changes behaviour from the start, this experiment explains how cooperation begins; identifying what triggers a later loss requires an intervention during learning.

F.8 Testing recovery through additional cooperative evidence

We test whether additional cooperative experience helps predominantly defecting learners return to cooperation. The protocol is fixed before generating outcomes. We run 256 independent intent-conditioned trajectories per noise $\epsilon = 0.175, 0.20, 0.225$ using generator seed 2026092603 and the horizon-two AllC configuration. After round 1,000, we select trajectories with intended cooperation at most 0.10 over rounds 901–1,000, giving 29, 139, and 251 selected runs. This targets current defection regardless of earlier behaviour. We copy each learner's complete state into three conditions: ordinary learning, 100 forced intended C actions with learning enabled, and the same forced actions with model updates withheld. All conditions resume ordinary action selection after round 1,100 and continue to round 5,000.

The two forced-cooperation conditions receive identical executed outcomes and retain state filtering using the forced actions. The no-update control discards parameter updates from the 100 probe transitions, including the final transition processed at round 1,101, while retaining all earlier evidence and subsequent ordinary learning. All three conditions share future execution-noise draws. This comparison separates the effect of updating the model from the experience of taking cooperative actions. Checks in four existing trajectories verified the update counts and confirmed that saving and restoring learner states reproduces uninterrupted trajectories.

The primary condition is $\epsilon = 0.20$. We compare final-quarter intended cooperation between the learning-enabled and no-update probes; the other noise levels are secondary. Paired 95% intervals use 5,000 bootstrap resamples of selected runs. All outcomes below exclude the forced actions, so they measure the learner's subsequent choices.

ϵ	n	Ordinary	Probe + learning	Probe, no updates	Difference, 95% interval
0.175	29	0.18171	0.29934	0.15305	0.14629 [0.00124, 0.29410]
0.20	139	0.02799	0.04388	0.02761	0.01626 [−0.01049, 0.04296]
0.225	251	0.00079	0.00011	0.00081	−0.00070 [−0.00093, −0.00048]

At the primary noise level, the cooperative prediction improves when the learner can update its model during the probe. On resuming ordinary action selection, the mean (CC, C) column concentration is 115.3 versus 32.0 without updates, and the predicted probability of joint CC is 0.820 versus 0.702. Both remain below the one-round reference $b^\star_\epsilon \approx 0.917$; the full planning decision also depends on continuation and information values. The interval for the late-cooperation difference nevertheless includes zero. During the first 500 rounds after the probe, cooperation is lower with learning (0.0272 versus 0.0503; paired difference −0.0230, interval $[-0.0389, -0.0076]$). At $\epsilon = 0.175$, learning improves late cooperation relative to the no-update probe, while the difference from ordinary learning remains uncertain (0.1176, interval $[-0.0300, 0.2561]$). At $\epsilon = 0.225$, all selected trajectories retain low late cooperation in every condition.

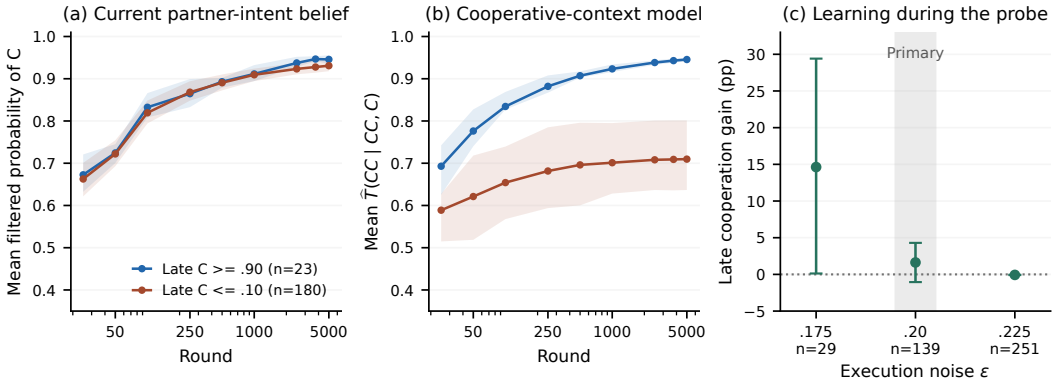


Figure 6: **Current intent beliefs, conditional learning, and a cooperation probe.** (a,b) Trajectories at $\epsilon = 0.20$, grouped retrospectively by final-quarter cooperation. Lines show group means and bands show across-seed interquartile ranges. The 53 intermediate trajectories are omitted from these panels. Similar beliefs about current partner intent coexist with different predictions of continued mutual cooperation. (c) Difference in final-quarter cooperation between the learning-enabled and no-update probes, in percentage points, with paired 95% bootstrap intervals. Counts give the runs selected using pre-intervention behaviour from 256 independently generated seeds per noise; $\epsilon = 0.20$ is the primary condition.

Additional evidence improves the cooperative prediction, but its effect on action choice also depends on the rest of the planning objective. In an exploratory comparison over rounds 1,101–1,600, the learning-enabled probe improves the mean payoff-only cooperation margin while leaving it negative, and the additional support from parameter-information value decreases. Each condition is evaluated under its own evolving beliefs. These observations suggest that learning changes both expected payoffs and information incentives, helping explain why a better cooperative prediction need not produce immediate behavioural recovery.

G Cross-game results

The cross-game distributions show how the same noise level can produce different behaviour depending on the payoffs, partner, and planner. Each game uses a 2×2 panel: rows show TFT and self-play; columns show intent-conditioned learners (BAPOMDP, AIF-POMDP) and observation-conditioned learners (Q-learning, AIF-MDP). The sweeps use 100 seeds, 5,000 rounds, and $\epsilon \in \{0, 0.05, \ldots, 0.25\}$, with settings in Appendix E. Points show per-seed final-quarter intended mutual cooperation, or intended agreement in Pure Coordination; lines and shading show medians and interquartile ranges.

The AIF pair isolates the effect of representation and channel placement under shared planning settings. BAPOMDP and Q-learning broaden the comparison to different learning and planning procedures. The captions identify the main behavioural differences, interpreted using the payoff comparisons in Appendix D.

G.1 Stag Hunt $(R, T, P, S) = (4, 3, 3, 0)$

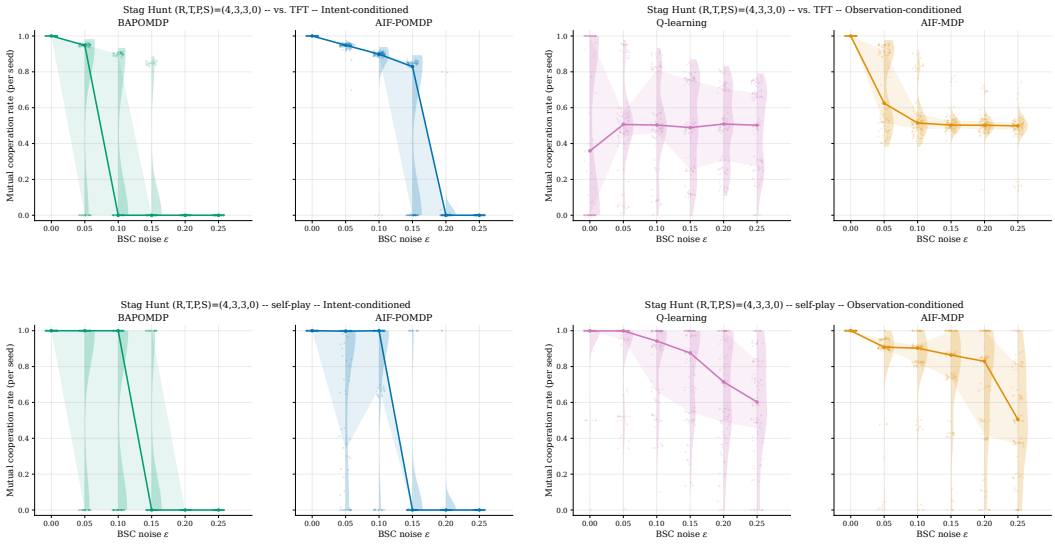


Figure 7: **Stag Hunt** $(R, T, P, S) = (4, 3, 3, 0)$ **across partner conditions.** Top row: TFT; bottom row: self-play. Left: BAPOMDP and AIF-POMDP; right: Q-learning and AIF-MDP. Points show per-seed intended mutual cooperation over the final quarter. Cooperation loss varies with both planner and partner: at $\epsilon = 0.15$, the mean rates are 0.178 (BAPOMDP) and 0.441 (AIF-POMDP) against TFT, and 0.260 and 0.110 in self-play.

G.2 Iterated Prisoner's Dilemma $(R, T, P, S) = (3, 5, 1, 0)$

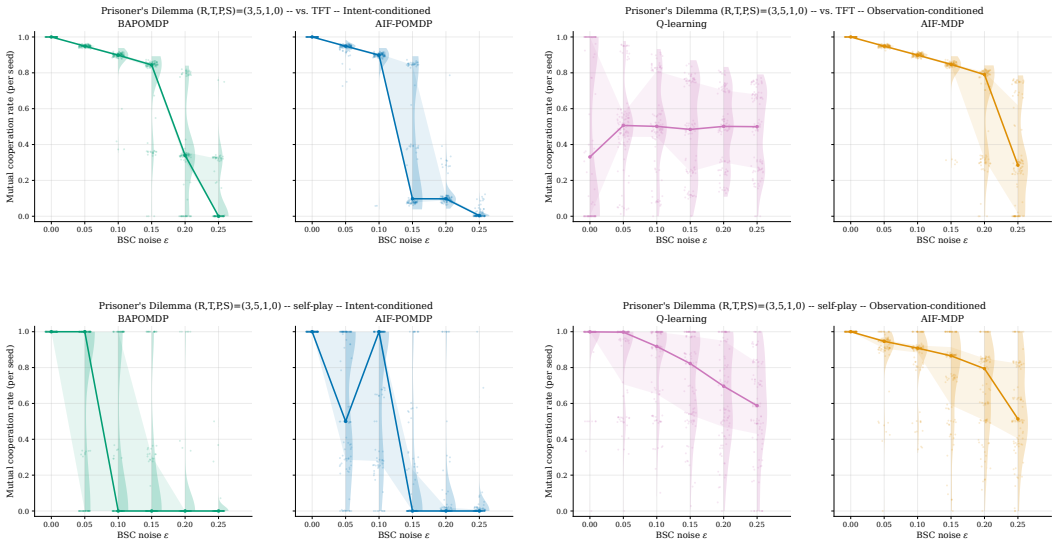


Figure 8: **Iterated Prisoner's Dilemma** $(R, T, P, S) = (3, 5, 1, 0)$ **across partner conditions.** Panel layout matches Figure 7; the metric is intended mutual cooperation. The intent-conditioned learners lose substantial cooperation as noise increases, with the extent of the loss depending on both planner and partner. At $\epsilon = 0.15$, BAPOMDP retains mean cooperation 0.734 against TFT and 0.212 in self-play; corresponding AIF-POMDP rates are 0.381 and 0.168.

G.3 CHICKEN $(R, T, P, S) = (3, 4, -1, 1)$

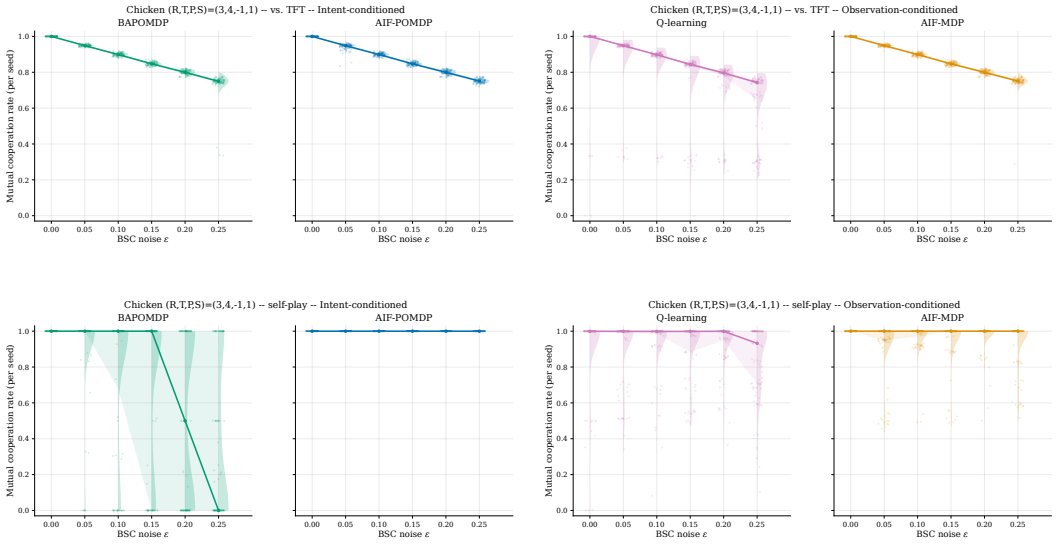


Figure 9: **Chicken** $(R, T, P, S) = (3, 4, -1, 1)$ **across partner conditions.** Panel layout matches Figure 7; the metric is intended mutual cooperation. In self-play, AIF-POMDP retains rate 1 across the tested noise grid, whereas BAPOMDP falls from 1 to 0.381 at $\epsilon = 0.25$. At this noise level, BAPOMDP's mean intended joint-action frequencies are 0.381 for CC, 0.351 for CD, 0.249 for DC, and 0.019 for DD, averaged over the same final-quarter windows and 100 seeds. Most of the reduction in mutual cooperation is therefore accompanied by asymmetric intended actions.

G.4 PURE COORDINATION $(R, T, P, S) = (1, 0, 1, 0)$

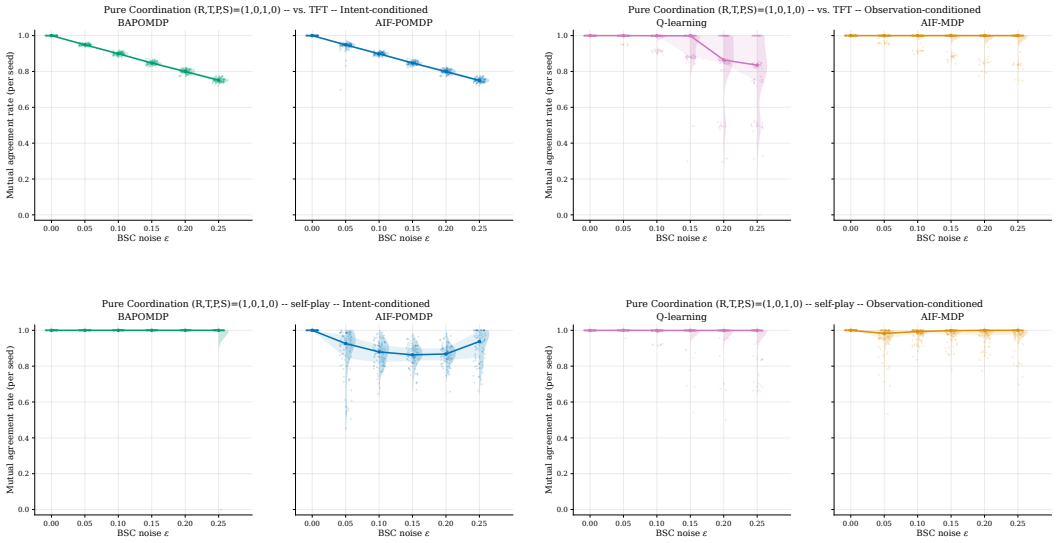


Figure 10: **Pure Coordination** $(R, T, P, S) = (1, 0, 1, 0)$ **across partner conditions.** Panel layout matches Figure 7; the plotted quantity is intended agreement on either C or D. At $\epsilon = 0.25$, mean intended agreement in self-play is 0.990 for BAPOMDP and 0.913 for AIF-POMDP; against TFT it is 0.751 for both. Corresponding mean realised agreement rates are approximately 0.625, 0.605, and 0.563. Equation equation 26 relates intended alignment to expected realised agreement under the execution channel.