THE DECISION VALUE OF PARTNER INFORMATION IN COOPERATIVE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative agents often need information that is available only to their partners. Auxiliary objectives that predict a teammate's behavior can improve internal representations, but prediction accuracy alone does not establish whether those representations improve coordination. We study this distinction by applying classical value of information to cooperative decisions and independently varying the availability and reward relevance of a partner's private information. Experiments in an exactly solvable commitment task and modified Cooperative Reaching environments separate the value of observing a partner from the benefit of training an agent to predict it. With memory trained through time, recurrent policies without prediction objectives recover the available history value in the toy task and match a scripted-partner reference in the continuously observed reaching task. Auxiliary prediction can substantially increase the decodability of a reward-irrelevant partner variable while leaving return equivalent within a prespecified margin. When partner observations are restricted to an initial window, however, some co-learning runs remain below a prespecified task-return threshold at the training budget. In this setting, a reward for behavior that makes the partner's goal inferable increases the fraction of training runs exceeding the threshold from 42/60 to 60/60 and improves mean task return by 0.131 (95% confidence interval [0.085, 0.181]), excluding the auxiliary reward; a matched reward for progress toward the goal does the same, and given four times the budget the unrewarded runs cross the threshold too, so the gain is one of speed. Teammate prediction objectives yield no statistically resolved improvement; an extra input branch read by the policy helps, whether or not it is trained to predict the partner. These results distinguish representational gains from behavioral gains and show that, under restricted observation, encouraging informative partner behavior can speed up coordination where prediction objectives on their own do not.

1 INTRODUCTION

When two agents cooperate and only one of them knows something important, such as where they should meet, the uninformed agent has to work it out tacitly from its partner's behaviour. A common way to help is to add an auxiliary objective that trains each agent to predict its teammate's next action, on the reasoning that an agent that models its partner should coordinate with it better (He et al., 2016; Hong et al., 2018; Hernandez-Leal et al., 2019; Papoudakis et al., 2021). These objectives are cheap to add and have been reported to raise return; in the cases we could check, the predicted variable was one the reward depended on, and none of these studies separately measured whether the objective changed the agent's representation, its behaviour, or both. Two observations from our experiments show why that distinction matters.

A learning agent plays beside a scripted partner that walks to a private corner. The learner can see the corner, but its own reward does not depend on it. Adding a teammate-prediction head raises the accuracy with which a linear probe reads the corner from the learner's memory from 0.42 to 0.98; control objectives with the partner information removed stay at 0.42 to 0.48. Return changes by at most 0.003 in every comparison (Section 6.2).

Now the partner's private corner is the only place where the pair gets paid, and both agents learn. Pairs trained with no prediction objective reach a return of 0.821, the same as a pair whose partner

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.

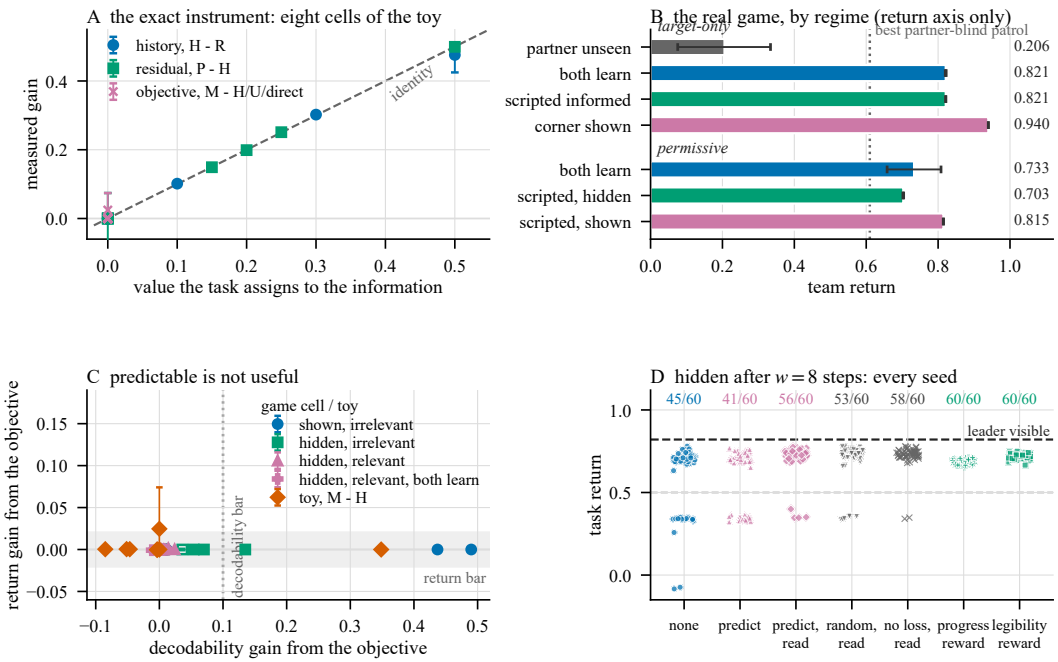

Figure 1: In our settings, return gains follow the value the task assigns to partner information, not how decodable the partner is. **(A)** Commitment task: the learner's gain lies on the identity line against the exact value. **(B)** Gridworld, return only: co-learning pairs reach the scripted-informed level. **(C)** The prediction objective raises decodability but no return difference is a benefit at the 0.02 margin; hollow marks are content-free controls. **(D)** Partner hidden after $w = 8$ steps, registered replication on fresh seeds: every seed's final return, with the count above the 0.5 threshold. Both leader rewards move seeds above it and the plain head does not; the branch the policy reads helps as much on random labels or with no prediction loss (grey) as with the prediction target (Section 7).

is scripted to walk straight to its corner (0.821), and far above a learner that cannot see its partner (0.21). Adding the prediction head changes return by -0.001 $[-0.003, +0.001]$ (Section 6.1). So making partner information decodable did not improve decisions, and where partner information did matter, agents obtained it without any modelling objective (Figure 1).

To make this precise we use the value of information (Blackwell, 1953; Raiffa & Schlaifer, 1961; Howard, 1966), applied to one cooperative decision: how much the agent's best decision improves if it learns the partner's private variable, given what it already observes. It is not a new quantity, and it depends on the task and the partner's policy together, not on the partner alone. In particular, a partner can be perfectly predictable and its information still worthless, if the predictable part never changes which action is best (Proposition 3). Predictability, decodability and mutual information are therefore different quantities from the value (Section 3).

We use two testbeds: a commitment task where the value has a closed form, so an agent's measured gain can be plotted against the exact value (Section 5); and a modified version of Cooperative Reaching (Rahman et al., 2023), in which one agent (the leader) has a private goal corner and the other (the follower) must find it. By changing the reward rule and what the follower can see, we create regimes where the partner's information is worth a lot, nothing, or something the agents only partly obtain (Section 6). The tested objective is always a head on the agent's recurrent network that predicts the partner's next action, compared against controls with the same parameters but no partner information; every agent's memory is trained through time. Table 1 summarises the regimes and what we found in each.

Without a reference for the decision value, a null return contrast is ambiguous: the variable may be worthless, already extracted by reward learning, unavailable from history, or represented but unused. Measuring the value first tells us which of these cases a result belongs to.

Table 1: The settings studied and the main result in each. "No objective" is a recurrent PPO agent without any prediction objective.

Setting	Is the partner's information worth anything?	What the no-objective agent does	Effect of the prediction head on return
Commitment task (Section 5)	Yes, computed exactly (0 to 0.5)	Obtains the computed value	None (equivalent in 7 of 8 cells)
Gridworld, only the leader's corner pays, both learn (6.1)	Yes (up to 0.33 of return)	Matches a scripted informed partner (0.821)	None: -0.001 $[-0.003, +0.001]$
Gridworld, corner irrelevant to the follower's reward (6.2)	No, by construction	Return unaffected	None, although decodability rises from 0.42 to 0.98
Gridworld, any shared corner pays, both learn (6.2)	No, given the convention the pair adopts	Pair fixes one meeting corner	None
Gridworld, scripted leader visible only for a window (6.3)	Yes, partly unrealised	Keeps the corner in memory but acts on it late	None
Gridworld, both learn, leader visible only for a window (7)	Yes, partly unrealised	Only 24 to 42 of 60 runs reach the threshold within the budget	Plain head: inconclusive. Head read by the policy: helps, and so does the same branch without prediction. Leader legibility reward: helps. No-objective pairs catch up at four times the budget.

We make four contributions. (1) *A task-level test before adding a partner model.* We operationalise the classical decision value of a partner's private variable as a task- and partner-policy-dependent reference. It separates the part of the variable recoverable from observation history from residual privileged information. Where the value is zero, revealing the variable cannot improve the best decision; where permitted history already supplies the attainable information, recurrent agents without an auxiliary objective recover it: exactly in the commitment task in seven of eight settings, with the remaining setting affected by one optimisation collapse, and up to the scripted-informed reference in the gridworld. (2) *A dissociation between representation and decision, and a boundary condition.* A plain next-action prediction head makes partner information far more decodable from memory, from 0.42 to 0.98 in the reward-irrelevant condition, while matched content-free controls remain at 0.42 to 0.48, without changing return. Three further prediction targets likewise yield no resolved return gain in the restricted-observation experiments. Giving the actor an extra input branch does help: a head whose features the actor reads improves finite-budget coordination, but in a fresh registered replication the same branch trained on random labels, or with no prediction loss at all, helps as much, and a LIAM-style encoder helps only once the task loss shapes it. In our experiments, behavioural gains came from features that the actor reads and the task loss shapes: the gain did not require partner prediction, and a parameter-matched wider network did not reproduce it. (3) *What helps under restricted observation, and what it buys.* When the partner is visible only briefly, leader-side shaping rewards for goal-directed movement and an extra actor-readable branch increase the fraction of successful runs. For the legibility reward, success rises from 42/60 to 60/60 at $w = 8$, and at $w = 5$ the effect persists with unshared parameters. With four times the training budget, no-objective agents reach the same coordination mode, 60/60 at $w = 8$ and 57/60 at $w = 5$, so the legibility reward, like the actor-readable branch, improves training speed rather than final policy quality in this task. (4) *An evaluation protocol for partner-modelling claims.* We combine task-computed information references, parameter-matched content-free controls, partner-input swaps that test causal use, and matched short- versus long-budget evaluations. This protocol separates changes in representation, behaviour, optimisation, and asymptotic policy quality that return or probe accuracy alone conflate (Section 7; step by step in Appendix G).

2 Related work

Predicting the other agent from a shared representation has repeatedly been reported to raise return (He et al., 2016; Hong et al., 2018; Hernandez-Leal et al., 2019; Papoudakis & Albrecht, 2020; Hernández et al., 2023; Raileanu et al., 2018; Huang et al., 2026). Papoudakis et al. (2021) train a separate encoder under partial observability against a fixed set of partners, and Huang et al. (2026) predict other agents' motion and rewards. In the papers we examined, the predicted variable matters to the reward by construction, and the representation is not measured separately from return. We add both measurements, with content-free controls and settings where the variable is irrelevant or already inferable. We do not claim that the published gains are wrong: they were obtained on other tasks with other learners, and we run their designs here as variants (Section 7).

The expected value of perfect information is due to Raiffa & Schlaifer (1961) and Howard (1966), and the comparison of information conditions to Blackwell (1953); our propositions (Section 3) are instances of these results. Closest in aim, Tessera et al. (2026) define diagnostics as expectations under the converged policy and, by their own scope statement, not as a task property independent of behaviour; their private-information measure asks by statistical dependence what our swap (Section 4) asks by intervention. Mon-Williams et al. (2025) read partner traits from recurrent states against a random-input floor, where we compare against a task-computed ceiling, and also find that such representations arise without auxiliary objectives. The convention gap of Fukushima et al. (2026) is a related restricted-information reference, and the partner swap they list as future work is our intervention. Cardei et al. (2026) share our motivation and define no value-of-information reference; Jing et al. (2026) find that supplying teammate actions gives inconsistent results, and Odrowaz-Sypniewski et al. (2026) that the best representation of intent is task-dependent. Our random-label controls play the role of the control tasks of Hewitt & Liang (2019).

Self-play conventions are arbitrary and can fail with new partners (Hu et al., 2021; 2020; Cui et al., 2021); which convention a group adopts is an equilibrium choice (Lerer & Peysakhovich, 2019); unequal payoffs, as in our permissive rule (Section 6.2), shape which convention forms (Sriraman et al., 2026); and Tessera et al. (2025) argue that benchmarks are often solved by conventions that bypass observation, as our permissive-rule pairs do. Work on signalling trains agents to act informatively because their actions will be interpreted (Foerster et al., 2019; Hu & Foerster, 2020): Strouse et al. (2018) regularise an informed agent to share or hide its goal; in a 5×5 grid a second agent takes about 30% longer to reach it when it is hidden; Jaques et al. (2019) reward causal influence on another agent's actions, Eccles et al. (2019) add biases for positive signalling and positive listening, Tian et al. (2020) reward helping the partner infer private state, Liu et al. (2025) add a legibility objective, Wang et al. (2025) reserve actions as a channel, and Grupen et al. (2022) measure emergent signalling. Our leader reward (Section 7) belongs to this family and is not new: it scores what Dragan et al. (2013) call legibility, motion that lets an observer quickly and confidently infer the correct goal, and is closest to the sharing regulariser of Strouse et al. (2018); the progress reward is distance-based shaping, which Ng et al. (1999) show can shorten learning. Our contribution is the comparison these rewards sit in, with prediction objectives of three published forms and a leader reward evaluated against the same reference and margins, each with matched controls.

Cooperative Reaching (Rahman et al., 2023; 2024) has no private information as published, so our variants are modifications. OvercookedV2 (Gessler et al., 2025) is used as published with a fixed partner (Appendix F); agents that assume an optimal partner can form protocols that fail with humans (Carroll et al., 2019). Our swap follows the advice of Lowe et al. (2019), and our intervals and inconclusive verdicts follow Agarwal et al. (2021).

3 The value of partner information

3.1 Definition

Let Z be the partner's private variable for the episode, H the information the focal agent holds when it decides, and $Q^\star(H, z, a)$ the value of action a when the partner's variable is z. The value of partner information at H is

$$\mathrm{VPI}(H) \;=\; \mathbb{E}_{Z|H}\Big[\max_a Q^\star(H, Z, a)\Big] \;-\; \max_a \mathbb{E}_{Z|H}\big[Q^\star(H, Z, a)\big] \;\geq\; 0, \tag{1}$$

the improvement in the agent's best decision from learning Z, given what H already says about it. The partner's policy and the dynamics are held fixed, and $Q^\star$ is the focal agent's optimal continuation value with the partner following its policy. The value is therefore a property of the task, the observation structure and the partner's policy together, not of Z alone. It is the expected value of perfect information (Raiffa & Schlaifer, 1961; Howard, 1966) written for one cooperative decision, and the ordering of information conditions below is Blackwell's (Blackwell, 1953).

We compare three information conditions: the reactive observation X at the decision, the permitted history $H \supseteq X$, and the partner's variable Z. Writing V_X, V_H and V_Z for the optimal expected reward under each, $V_X \leq V_H \leq V_Z$. The total value $V_Z - V_X$ splits into the *history value* $V_H - V_X$, which memory of the partner's behaviour can obtain, and the *residual* $V_Z - V_H$, which no policy on the permitted history can obtain. For a single decision the residual equals $\mathbb{E}_H[\mathrm{VPI}(H)]$; in continued play the ordering still holds but the residual is only bounded. A prediction objective that sees the same history as the learner adds no information. It can only change what the learner computes from that history.

Appendix D proves three consequences that we use throughout. The value is zero exactly when one action is optimal for every partner variable the posterior supports (Proposition 1). It equals the probability that the variable changes the optimal action times the expected gap in action value, so it can be computed from a posterior (Proposition 2). Evidence about a part of the variable that the value does not depend on leaves the value unchanged, while the mutual information can grow without bound (Proposition 3).

3.2 Two instances, never plotted on one axis

The *decision instance* scores a single declared corner: $Q^\star$ is 1 when the follower's declared corner matches the partner's. When only the partner's corner pays, Proposition 2 gives $\mathrm{VPI}(H) = 1 - \max_z P(Z = z \mid H)$. Against a scripted partner with a known movement rule the posterior is exact, so this value is computed, not estimated. The accuracy of the best single declaration rises from 0.55 after one observed move to 0.94 after eight (Monte Carlo half-width 0.022), and for a follower that sees the partner only for its first w moves it is 0.63, 0.79 and 0.91 at $w = 3, 5, 8$. This bounds one irrevocable declaration, not return over an episode. The *return instance* uses episode return and includes timing and execution, so it is only bounded. In the game where only the partner's corner pays, an oracle that sees the corner earns 0.94 and the best partner-blind scripted policy we found earns 0.61, so partner information is worth at most 0.33 of return.

For a recurrent learner, a shortfall against the value can come from three places: the history never contained the information, the memory did not keep it, or the policy kept it but did not use it. Only the first is computed; the other two are bounded from one side each. We checked both instances on scripted followers with known internal states before applying them to learned agents (Appendices D.3 and D.1). There, decodability and mutual information measure what is present, return also reflects execution, and neither is the value: mutual information ranks one case above another, and return ranks a second pair, in the opposite order to the value (Table 2).

4 Agents, objective and statistics

Every agent is a recurrent PPO policy (Schulman et al., 2017) whose memory is trained through time; an attention-over-history variant is used where stated. Both were checked on a delayed-recall task before use. When both agents learn, one network controls both (parameter-shared self-play) unless unshared parameters are stated. Every reported return is the training return over the last tenth of 400 iterations of 256 episodes each, or of 1600 iterations where stated (Appendix B).

The objective is a head on the recurrent network that predicts the partner's next action, at weight 0.05 (and 0.2 in one experiment), on either agent or both. Every configuration (*arm*) carries the same head; the no-objective arm trains it at weight zero, so all arms have the same parameters. Three controls keep the head and its gradient but remove the partner content: random labels, the agent's own next action, and the head with its gradient stopped at the network. Privileged diagnostics, which give the head the partner's variable, are used in training only and labelled as such.

Decodability is the held-out accuracy of a linear probe that reads the partner's variable from the agent's memory. *Legibility* $L(k)$ is the same accuracy for a probe that reads the informed agent's goal from its first k moves. To test whether a trained agent *uses* its partner, we replace the partner's inputs at evaluation with those of another episode (the *swap*), which keeps their distribution but removes their information; replacing them with zeros would do neither (Lowe et al., 2019).

Seeds are fixed before each run and none is dropped: twenty per arm in the toy, ten in the game with a scripted or always-visible partner, and sixty when the partner is hidden after a window and both agents learn. Each comparison is an unpaired mean difference with a seed-level bootstrap 95% interval, a permutation test (Monte Carlo where the relabellings cannot all be enumerated), and Holm correction within a family declared in advance. We call an effect a *benefit* when the interval lies above a practical margin (0.02 of return, 0.10 of decodability), *equivalent* when it lies inside the margin, and *inconclusive* otherwise; equivalent is a practical judgement, not a claim of exact equality. When the partner is hidden after a window and both agents learn, outcomes are bimodal, so the primary outcome is the fraction of runs whose final return exceeds 0.5, with exact binomial intervals and Fisher tests. Its threshold and its 0.2 margin were fixed before the runs: a rate gain counts when the Fisher test gives $p < .05$ and the observed difference is at least 0.2. The counting rule was corrected once, before the second window was run (Appendix H.2). The critic is centralised and reads both agents' observations, so in the gridworld it sees the leader's corner during training; no policy does, and every arm uses the same critic. A consistency check requires each update to reproduce the action probabilities of the rollout (largest mismatch 5.7×10^{-6}).

5 A COMMITMENT TASK WITH AN EXACT VALUE

We first test the value where it has a closed form. In a short commitment task a type Z is drawn uniformly, a cue that matches it with probability p is shown once, and after five blank steps the agent commits; a safe action pays $1 - g$ and an action that complements the type pays g. The history value is then $\max(0,\ gp - \frac{1}{2})$. Eight settings of (g, p) place it at 0, 0.1, 0.3 and 0.5, with twenty seeds per arm.

A recurrent learner with no objective gains $+0.101$, $+0.302$ and $+0.476$ where the value is 0.1, 0.3 and 0.5, and stays within 0.005 of zero in the five settings where the value is zero (Figure 1A). One of those five has a cue that predicts the type 65% of the time but never changes the best action. The prediction head is equivalent to the no-objective learner in seven of eight settings; the eighth is inconclusive because one seed collapsed. A network with 2.5 times the parameters changes nothing, so the head's null is not a capacity effect. Where the type is perfectly predictable but worth nothing, any auxiliary head raises its decodability from 0.65 to 1.00 with a return change of 0.000, as Proposition 3 implies (Appendix A).

6 COOPERATIVE REACHING: THE VALUE BY REGIME

Cooperative Reaching (Rahman et al., 2023) is a 5×5 grid in which two agents are paid when they stand together on a corner. The published game carries no private information, so we modify it. The leader is given a private goal corner and waits for a private delay of 0 to 6 steps; each of its actions is then replaced by a random one with probability 0.4. The follower pays 0.02 per step away from a rewarded corner, so waiting has a cost. Both agents see both positions and the partner's previous action; in the hidden condition the follower never sees the goal corner. Under the *target-only* rule only the leader's corner pays. Under the *permissive* rule any shared corner pays (1.0 at two corners, 0.75 at the other two). A scripted leader walks by a known rule (Appendix B), so the follower's posterior is exact.

6.1 WHEN THE PARTNER'S CORNER IS THE ONLY ONE THAT PAYS

Both agents learn under the target-only rule, with recurrent (GRU) and attention learners and ten seeds per arm (Table 4). All 140 runs converged. Pairs with no objective reach 0.821 (GRU) and 0.815 (attention), the same as a pair whose leader is scripted to walk to its corner (0.821 and 0.818). Showing the follower the corner raises return to 0.940, and a follower that cannot see its partner reaches only 0.21 ± 0.22. When the leader's inputs are swapped with another episode's at

evaluation, every hidden-corner pair drops to between -0.32 and -0.41, while pairs that are shown the corner are unaffected; the follower therefore uses the leader. The prediction head changes return by -0.001 $[-0.003, +0.001]$ with the GRU and $+0.002$ $[-0.001, +0.004]$ with attention. Every placement of the head, the random-label control and the unshared-parameter arms are equivalent to no objective. In this regime, then, co-learning agents read the partner's behaviour well enough to match the scripted reference, and the head adds nothing measurable. (These numbers come from a repetition on a second platform; the original run agrees within 0.003 and on every verdict, Appendix H.1.)

6.2 When the corner is irrelevant to the follower's reward

Against the scripted leader, a 2×2 design crosses whether the corner is shown or hidden with whether it is relevant or irrelevant to the follower's reward, with ten seeds per arm (Table 5). Return does not move: all twenty contrasts against no objective lie within 0.003, and the 0.111 cost of hiding a relevant corner is unchanged by every objective and control. The representation does move. When the corner is shown but irrelevant, the head raises its decodability one step after the partner is seen from 0.42 to 0.96 and 0.98 at the two weights, while the stop-gradient and random-label controls stay at 0.42 to 0.48. The commit step and return do not change. When the corner is hidden and relevant, no arm changes the memory's content at all: with memory trained through time the learner already holds what the objective could add. These results match Propositions 1 and 3: the head changes what is linearly decodable from memory, and in this design that change does not reach the decision. The same dissociation appears in OvercookedV2: the head raises the recipe's decodability by $+0.13$ and predicting the recipe itself by $+0.43$, while no arm changes return by the margin (Appendix F).

Under the permissive rule, where any shared corner pays, all 140 co-learning pairs instead settle on one meeting corner. That corner agrees with the leader's private corner only at chance, the leader's route carries no information about it ($L(3) = 0.25$ against 0.65 for the scripted leader), and no placement of the head changes this. Given the convention each pair adopts, the partner's information is worth nothing, in line with prior findings on self-play conventions (Hu et al., 2021; Tessera et al., 2025) (Appendix E).

6.3 When the leader is visible only for a window

In the regimes above the no-objective learner already reaches its reference. To find a regime where it does not, the follower now sees the scripted leader only for its first w moves after the delay and must then act on what it remembers. The decision value at the end of the window is 0.63, 0.79 and 0.91 at $w = 3, 5, 8$. We registered this experiment in advance as a diagnosis, with a table mapping each possible outcome to one reading (Appendix D.2; arms in Table 6).

The windowed follower earns 0.708, 0.752 and 0.781 against 0.8215 with the leader always visible. It loses 0.20 to 0.60 under the swap and sits 0.12 to 0.20 above followers that never see the leader, so the lost return is tied to the hidden signal.

When the window closes, the corner is decodable from memory at 0.58, 0.74 and 0.79, within 0.03 to 0.07 of the best achievable on the same episodes. Yet the follower commits to a corner 8.5 and 4.4 steps after the window closes at $w = 3$ and 5, and at the close at $w = 8$.

The prediction head does not change this: every return contrast lies within $[-0.004, +0.007]$ and every decodability contrast at the close is inside the 0.10 margin; later decodability and commit timing are inconclusive at ten seeds (Table 6). We had expected the head to improve retention, because the hidden leader's next action is predictable only from the remembered corner. It did not: memory, commitment and return match the no-objective follower's. This describes where the information sits, not why return is lost; Section 7 tests one causal reading of it.

7 Where agents fall short: what helps and what does not

The windowed regime is the one place where agents without an objective fall short. We test changes on both sides of the pair: the follower's decision and representation, and the leader's behaviour.

Against the scripted leader, pulling the follower toward the corner its head decodes is equivalent to no change at weight 0.05 and harmful at 0.2, and a follower that simply follows its most likely corner earns less than the unmodified one (Appendix C.1). Holding the information is therefore not enough; how the policy uses it matters.

All remaining experiments train both agents, with a window of $w = 8$ and a replication at $w = 5$, and 60 seeds per arm. Here outcomes are bimodal: a run either learns to use the leader or settles on a fixed route over two or three corners that is paid only when the goal lies on it (Appendix I). We therefore count the runs above a return of 0.5. With no objective, 42 of 60 runs reach it at $w = 8$ and 24 of 60 at $w = 5$; repeated no-objective arms vary by about 8 of 60 between experiments, so each comparison below uses its own no-objective arm.

We add a shared reward on the leader's movement, the *legibility reward*, paid when the window closes, of $\beta\,(P_{\mathrm{obs}}(z) - 1/4)$, where $P_{\mathrm{obs}}(z)$ is the probability an observer assigns to the leader's true corner under the scripted leader's movement rule and $1/4$ is chance ($\beta = 0.5$ and 2.0). The follower's inputs and objective are unchanged. At $w = 8$ all 60 of 60 runs reach the threshold at both weights (Fisher $p = 1.7 \times 10^{-6}$), and mean task return, excluding the reward, rises by 0.131 $[0.085, 0.181]$ and 0.083 $[0.038, 0.133]$ (Table 8). At $w = 5$, 58 and 60 of 60 runs reach it, and return rises by 0.260 $[0.203, 0.317]$. The same reward drawn from another episode leaves 44 of 60, and a reward for revealing a corner the task never uses leaves 47 of 60, so neither reward noise nor arbitrary shaping explains the gain. Among runs that reach the threshold, the lower weight changes return by only $+0.013$, and at the higher weight every run ends between 0.647 and 0.691: the reward removes low runs rather than raising good ones.

A control registered in advance pays instead for the fraction of the leader's distance to its goal covered within the window, with no observer involved. It works as well, 57 and 60 of 60 runs against 32 in its own experiment (return $+0.165$ $[0.104, 0.224]$; Table 7), and raises the leader's legibility about as much. In this game the fastest route to a corner also reveals it, so the two rewards cannot be separated, and we do not claim a signalling effect. Consistent with this, the legibility probe misses its registered 0.10 margin at both windows ($+0.045$ to $+0.123$), while the follower decodes the corner better at the window's close ($+0.157$ $[0.134, 0.179]$) and commits 3.8 steps earlier, at $w = 5$ (Table 7).

The head on the follower, on the leader, and on the leader with random labels reaches the threshold in 36, 44 and 40 of 60 runs against 42, with return contrasts from -0.026 to $+0.006$ and intervals inside $[-0.09, +0.08]$ (Table 8). These results are inconclusive at this sample size, not evidence of no effect. This plain head corresponds to the AMS design of Hernandez-Leal et al. (2019); we also test their AMF design, in which the head's hidden features are fed to the policy, with and without the head's gradient reaching the network. At $w = 8$, in a separate experiment with its own no-objective arm at 35 of 60, the two feature-reading variants improve return by 0.167 $[0.113, 0.221]$ and 0.134 $[0.076, 0.194]$, with 58 and 53 of 60 runs above the threshold, while the plain head is again inconclusive ($+0.034$ $[-0.034, +0.103]$; Table 9). At $w = 5$ the feature-reading variants help less ($+0.122$ and $+0.100$, 46 and 42 of 60 against 32), and one misses its registered rate margin (Table 7). Here the gain is on the follower's side: it decodes the corner better at the close ($+0.078$) and commits 2.6 steps earlier, while the leader's route does not change. A fresh replication registered in advance (unused seeds, one code version, 420 runs; Table 11) reproduces the ordering: 45 of 60 runs reach the threshold with no objective, 41 with the plain head (-0.006 $[-0.072, +0.062]$), 56 with the reading head ($+0.105$ $[+0.052, +0.164]$) and 60 with either reward ($+0.083$ and $+0.115$). It also includes two controls fixed in advance: the same reading branch trained on random labels reaches 53 of 60 ($+0.086$ $[+0.028, +0.148]$), and the branch with no prediction loss reaches 58 of 60 ($+0.118$ $[+0.067, +0.175]$). The registered content contrast, the reading head minus the same branch on random labels, is $+0.018$ $[-0.021, +0.061]$, and the reading head minus the loss-free branch is -0.013 $[-0.044, +0.017]$. By the rule fixed before the run, the experiment does not support a contribution from predicting the partner: the gain comes from letting the policy read a nonlinear layer of its memory that the task loss shapes. A second registered control on fresh seeds rules out parameter count: a no-objective trunk widened to the branch's parameter count does not resolve a gain ($+0.036$ $[-0.039, +0.108]$), while the loss-free branch is a benefit over it and over no objective ($+0.127$ $[+0.074, +0.185]$ and $+0.163$ $[+0.111, +0.216]$; Table 12).

Changing what the head predicts on the shared network does not resolve an effect at $w = 5$: the next action ($+0.030$ $[-0.044, +0.103]$), the partner's next three positions as in Huang et al. (2026)

(-0.011 and $+0.015$ at two weights), and even the private corner itself as a privileged label ($+0.054$ $[-0.020, +0.130]$). A LIAM-style arm (Papoudakis et al., 2021), a separate encoder trained by reconstruction whose embedding is passed to the policy without gradient, is also inconclusive at both windows ($+0.045$ $[-0.018, +0.105]$ and $+0.077$ $[+0.001, +0.149]$; Table 9). When that gradient block is removed, the encoder helps at $w = 8$ ($+0.098$ $[+0.032, +0.163]$, 47 against 36 of 60; Fisher $p = 0.047$), although the rate gain of $+0.18$ misses its registered 0.2 margin. Across these variants, every gain we observed came from features that the actor reads and the task loss shapes, and the replication shows that a prediction target is not needed for it.

Two further registered experiments test the reward's robustness. Trained for four times as long, the no-objective pairs also reach the threshold, in 57 of 60 runs at $w = 5$ and 60 of 60 at $w = 8$, and the reward's return advantage shrinks to $+0.003$ $[-0.017, +0.028]$ and -0.038 $[-0.043, -0.033]$ (Table 10, Figure 2). The low runs are therefore slow, not stuck, and the reward makes the pair reach the informed behaviour sooner rather than better. Rewarded leaders are also easier for separately trained followers: followers trained against frozen rewarded leaders do 0.18 to 0.19 better with held-out leaders than followers trained against unrewarded ones, which reflects both more reliable rewarded leaders and better transfer across leaders (Table 13). With one network per agent at $w = 5$, both rewards still bring 60 of 60 runs above the threshold against 40 (return $+0.160$ and $+0.165$), and the feature-reading head brings 54 of 60 ($+0.110$). Trained for four times as long at $w = 8$, the feature-reading head is equivalent to its own no-objective pair ($+0.005$ $[+0.000, +0.010]$, both 60 of 60), so its gain, like the reward's, is one of speed.

8 Limitations

We study two agents with one private variable per episode, model-free recurrent PPO learners, and modified Cooperative Reaching, with OvercookedV2 as a second setting (Appendix F). The main prediction method is a next-action head at weight 0.05; the AMS, AMF, LIAM-style and motion-target designs are run as variants of it (Appendix B), not as full reimplementations. Partners whose identity changes the best response in ways not readable from their route are untested. Our results do not show that teammate modelling never helps; they show when it did not in these settings.

The propositions are instances of classical results, and the value must be written for each task, so it is a way of describing coordination problems rather than a general metric. Computing it before adding an objective requires a model of the partner's behaviour. The per-window values bound a single declaration, not return over an episode, and matching a scripted informed partner means matching a reference, not an optimum: pairs shown the corner still do better.

We do not decompose the gap to the oracle in the positive regime. The plain prediction head and the other shared-network targets are inconclusive at 60 seeds, not shown to be ineffective. The leader reward's gain is one of speed, and at four times the budget it gives no advantage at $w = 5$ and a small disadvantage at $w = 8$, and the feature-reading head's gain is also one of speed. The reward is not evidence for signalling, since a reward for distance covered works as well. Zero failures in 60 seeds still allows a failure rate of up to about 6%. On OvercookedV2 no objective, and no extra branch read by the policy, changes return by the margin, and two arms remain inconclusive at twenty seeds; the finding that such a branch helps is stated for the gridworld only. Two of forty no-objective toy runs collapsed after reaching the optimum. Reproduction coverage and every departure from the registered plans are in Appendices H.1 and H.2.

9 Conclusion

We measured the value of a partner's private information for each task and compared it with what learning agents obtain. In our settings, recurrent agents without a prediction objective reached the exact value in a commitment task and the level of a scripted informed partner in Cooperative Reaching, and a prediction head changed what their memory contained without changing return. Where agents fell short, a reward on the informed agent's goal-directed movement made training reach the informed behaviour faster, and so did an extra input branch read by the policy, whether or not it predicted the partner. Before adding a teammate-prediction objective, it is worth asking whether the partner's information changes any decision, and whether the agent already obtains it.

REPRODUCIBILITY STATEMENT

For every experiment, the arms, seeds, margins and the reading of each outcome were fixed in its configuration before it was run, and every reported number is computed by a single statistics module from the per-run results. Appendix H.1 reports which experiments were repeated and how closely the repetitions agree. The environments, the objective, the swap test and the legibility reward are specified in Appendices A and B. The code, configurations, per-run results and the scripts that produce every table and figure will be released with the paper.

STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

The authors used a large language model as a coding and writing assistant: to implement experiment code, to run the pre-registered analyses, and to revise manuscript text. Every experimental design, code, decision, pre-registration and claim in this paper was made or checked by the authors, who take full responsibility for the content.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Maria Ana Cardei, Matthew Landers, and Afsaneh Doryab. Coordination matters: Evaluation of cooperative Multi-Agent reinforcement learning, 2026. URL https://arxiv.org/abs/2605.06557.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for Human-AI coordination. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob N. Foerster. K-level reasoning for zero-shot coordination in Hanabi. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8215–8228, 2021.

Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 301–308, 2013. doi: 10.1109/HRI.2013.6483603.

Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. Biases for emergent communication in multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13111–13121, 2019.

Jakob 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. In *International Conference on Machine Learning (ICML)*, pp. 1942–1951, 2019. URL https://proceedings.mlr.press/v97/foerster19a.html.

Makoto Fukushima, Hua-Dong Xiong, and Ehsan Moradi Pari. The convention gap: Towards measuring implicit communication in cooperative AI evaluation, 2026. URL https://arxiv.org/abs/2609.11489.

Tobias Gessler, Tin Dizdarevic, Ani Calinescu, Benjamin Ellis, Andrei Lupu, and Jakob Nicolaus Foerster. OvercookedV2: Rethinking Overcooked for zero-shot coordination. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=hlvLM3GX8R.

Niko A. Grupen, Daniel D. Lee, and Bart Selman. Multi-agent curricula and emergent implicit signaling. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 553–561, 2022. doi: 10.65109/pryn2201.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. Opponent modeling in deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 1804–1813, 2016. URL https://proceedings.mlr.press/v48/he16.html.

Daniel Hernández, Hendrik Baier, and Michael Kaisers. BRExIt: On opponent modelling in expert iteration. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3795–3802, 2023. doi: 10.24963/ijcai.2023/422.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. Agent modeling as auxiliary task for deep reinforcement learning. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2019.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Zhang-Wei Hong, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. A deep policy inference Q-network for multi-agent systems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 1388–1396, 2018. doi: 10.65109/tjsn8981.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Hengyuan Hu and Jakob N. Foerster. Simplified action decoder for deep multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=B1xm3RVtwB.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-Play" for zero-shot coordination. In *International Conference on Machine Learning (ICML)*, pp. 4399–4410, 2020. URL https://proceedings.mlr.press/v119/hu20a.html.

Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning (ICML)*, pp. 4369–4379, 2021. URL https://proceedings.mlr.press/v139/hu21c.html.

Qinwei Huang, Simon Khan, Rui Zuo, Stefan Wang, Garrett E. Katz, and Qinru Qiu. Predictive auxiliary learning for belief-based Multi-Agent systems. In *Proceedings of the AAAI Symposium Series*, volume 8, pp. 50–58, 2026. doi: 10.1609/aaaiss.v8i1.42517.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049, 2019.

Yuheng Jing, Kai Li, Ziwen Zhang, Jiajun Zhang, Zeyao Ma, Jiaxi Yang, Lei Zhang, Zhe Wu, Jinmin He, Junliang Xing, and Jian Cheng. Benchmarking the limits of In-Context reinforcement learning for Ad-Hoc teamwork. In *International Conference on Machine Learning (ICML)*, 2026.

Adam Lerer and Alexander Peysakhovich. Learning existing social conventions via observationally augmented self-play. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2019.

Yanyu Liu, Yinghui Pan, Yifeng Zeng, Biyang Ma, and Prashant Doshi. Active legibility in multiagent reinforcement learning. *Artificial Intelligence*, 346:104357, 2025. doi: 10.1016/j.artint.2025.104357.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

Ruaridh Mon-Williams, Max Taylor-Davies, Elizabeth Mieczkowski, Natalia Vélez, Neil R. Bramley, Yanwei Wang, Thomas L. Griffiths, and Christopher G. Lucas. Partner modelling emerges in recurrent agents (But Only When It Matters). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287. Morgan Kaufmann, 1999.

Mateusz Odrowaz-Sypniewski, Jasmine Bayrooti, Ajay Shankar, and Amanda Prorok. Generalized intention modeling in Multi-Agent reinforcement learning, 2026. URL https://arxiv.org/abs/2605.31318.

Georgios Papoudakis and Stefano V. Albrecht. Variational autoencoders for opponent modeling in multi-agent systems. In *AAAI-20 Workshop on Reinforcement Learning in Games*, 2020.

Georgios Papoudakis, Filippos Christianos, and Stefano V. Albrecht. Agent modelling under partial observability for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Arrasy Rahman, Elliot Fosong, Ignacio Carlucho, and Stefano V. Albrecht. Generating teammates for training robust ad hoc teamwork agents via best-response diversity. *Transactions on Machine Learning Research*, 2023.

Arrasy Rahman, Jiaxun Cui, and Peter Stone. Minimum coverage sets for training robust ad hoc teamwork agents. In *AAAI Conference on Artificial Intelligence*, 2024.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. Modeling others using oneself in multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 4257–4266, 2018. URL https://proceedings.mlr.press/v80/raileanu18a.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Abhishek Sriraman, Eleni Vasilaki, and Robert Loftin. ConventionPlay: Capability-limited training for robust ad-hoc collaboration, 2026. URL https://arxiv.org/abs/2604.18123.

DJ Strouse, Max Kleiman-Weiner, Josh Tenenbaum, Matt Botvinick, and David Schwab. Learning to share and hide intentions using information regularization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10270–10281, 2018.

Kale-ab Abebe Tessera, Leonard Hinckeldey, Riccardo Zamboni, David Abel, and Amos Storkey. Remembering the Markov property in cooperative MARL. In *Reinforcement Learning Conference (RLC) Workshop on Finding the Frame*, 2025.

Kale-ab Abebe Tessera, Leonard Hinckeldey, Riccardo Zamboni, David Abel, and Amos Storkey. Probing Dec-POMDP reasoning in cooperative MARL. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026. doi: 10.65109/ECCJ1033.

Zheng Tian, Shihao Zou, Ian Davies, Tim Warr, Lisheng Wu, Haitham Bou Ammar, and Jun Wang. Learning to communicate implicitly by actions. In *AAAI Conference on Artificial Intelligence*, 2020.

Han Wang, Binbin Chen, Tieying Zhang, and Baoxiang Wang. Learning to communicate through implicit communication channels. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=wm5wwAdiEt.

A THE COMMITMENT TOY IN FULL

The toy is a measuring instrument with closed forms, not a game. Its partner does nothing until the commitment step and then plays its type, so no claim about two learning agents rests on the toy. A type $Z \in \{L, R\}$ is drawn uniformly. At step 0 the agent sees a cue C with $P(C = Z) = p$. It also sees a second cue, of equal reliability, about an attribute that never enters the reward. After five

blank steps the agent commits to an action a and receives $(1-g)\,\mathbf{1}[a = L] + g\,\mathbf{1}[a \neq Z]$. The safe action is therefore worth $1-g$, and the action that depends on the type is worth g. The exact values are $V_X = 1 - g/2$, $V_H - V_X = \max(0,\, gp - \frac{1}{2})$ and $V_Z - V_X = \max(0,\, g - \frac{1}{2})$. Here V_X is the value of a policy that ignores the cue, V_H the value with the history, and V_Z the value with the type shown; we call $V_Z - V_H$ the residual. In words, g sets whether the type matters, and p sets whether the history carries it.

Six arms (configurations of the experiment) train on the same seeds:

- R, a reactive learner that does not see the cue;
- H, the history learner with no objective;
- M, the next-action head, whose target is a no-op before the commitment step and the type at that step;
- U, the same head predicting the irrelevant attribute;
- M-direct, a head whose target is the type itself;
- P, a learner that is shown the type.

Eight settings of (g, p) place the value of history at 0, 0.1, 0.3 and 0.5 and the residual between 0 and 0.5. Each arm has 20 seeds per setting. We registered four falsifiers in advance. Cue leakage: R scores above its exact value. Type leakage: an arm without the type scores above V_H. Unused privilege: a privileged input is not used. Memory or optimisation failure: H scores below 0.98 where $V_H = 1$.

Table 3 places every trained arm against its closed form. Where the task puts the value of history at 0.1, 0.3 and 0.5, the history learner gains $+0.101$ $[+0.099, +0.104]$, $+0.302$ $[+0.298, +0.306]$ and $+0.476$ $[+0.425, +0.503]$ over the reactive learner. In the five settings where that value is zero, its gain is within 0.005 of zero. Where the residual is 0.25, 0.15, 0.2 and 0, the privileged arm gains $+0.251$, $+0.149$, $+0.199$ and $+0.000$ over the history learner. Each of these intervals lies within 0.004 of the residual. In the boundary settings every arm is within 0.0023 of its exact value, and neither leakage falsifier triggered. One falsifier did trigger: the memory or optimisation check. At $g = 1$, $p = 1$, one of twenty H seeds and one of twenty P seeds reached the ceiling and then fell to 0.5 within a single policy update. This pulls the H mean down to 0.975, and the gain to 0.476 where the exact value is 0.5. The medians sit at 1.000, as does a follow-up with 100 fresh seeds per arm. Every collapsed run had first peaked at 0.999 or above. This is consistent with the memory having been learned and the failure being one of optimisation. In the follow-up, H, P, M and U each had 0 collapses in 100 seeds. An arm trained on unlearnable noise had 1 (Fisher $p = 1.0$ and 0.33 against the pooled no-objective arms). The two collapses in forty runs of the core experiment had a chance probability of 0.16, so they are consistent with chance. We make no claim about stability in either direction.

In the setting $g = 0.75$, $p = 0.65$ the cue predicts the type with accuracy 0.65 but changes no decision. There the history learner gains -0.0003 $[-0.0036, +0.0029]$ over the reactive one, while a probe reads the type from its memory at 0.63 at step 0. In the settings with $g = 0$ the cue predicts the type perfectly, but the type is worth nothing, and every arm earns 1.000. In these settings the next-action head keeps the worthless type decodable from memory at 1.00 throughout the delay. This is $+0.35$ $[+0.29, +0.40]$ above the no-objective learner's mean over steps of 0.65, with a return gain of 0.000.

M is equivalent to H, within the 0.02 margin registered in advance, in seven of eight settings. In the eighth ($g = 1$, $p = 1$) the result is inconclusive, $[-0.0002, +0.0741]$, because of the collapsed H seed. Over the 200 fresh no-objective seeds of the follow-up the difference is 0.000; this is a descriptive reading, not a registered test. M is equivalent to U and to M-direct in all eight settings. In the boundary settings the largest of the twelve objective contrasts is 0.0012. A no-objective network with 2.5 times the parameters ($355{,}844$ against $142{,}084$) is equivalent to the small one and to M in the two settings tested. The head's effect, where it has one, is therefore not a capacity effect. In the three boundary settings where the cue is imperfect ($p < 1$), the next-action head lowers type decodability by 0.05 to 0.09, with return equivalent. Its target depends on the type only at the commitment step, and in these runs a head trained on that target does not keep the cue in memory. The toy thus measures the value side of the paper. A learner with no modelling objective

collects the value of partner history where the task puts it, and nowhere else. When information is matched, the objective is equivalent to no objective wherever the test resolved. The toy leaves no gap in decodability for the objective to close, so the dissociation between representation and return rests on the game.

B LEARNERS, OBJECTIVE, CONTROLS AND STATISTICS

Every agent is a recurrent PPO policy. A GRU keeps a belief over the observation history and feeds an actor and a critic. The critic is centralised: it reads the concatenated observations of both agents. In the gridworld the leader's observation holds its own corner, so the critic sees the corner during training, in every arm including no objective; the follower's policy never does. In the commitment task the partner's observation is empty, so the critic sees only what the learner sees. Training carries the gradient through the whole episode (sequence training). An attention-over-history variant replaces the GRU where stated. Both were checked on a delayed-recall task before use. An *experiment* is a set of arms trained together on the same seeds.

Every run is 400 iterations (1600 in the two budget experiments below). Each iteration collects 256 parallel episodes of horizon 50. The remaining settings are:

- PPO: ten epochs per iteration on that iteration's batch, clip 0.2, entropy coefficient 0.05.
- Returns: $\gamma = 0.99$ and GAE $\lambda = 0.95$.
- Optimiser: Adam at 5×10^{-4} for actor and critic, with linear annealing, and gradient clip 10.
- Network: a GRU of 128 units over a 64-unit observation embedding, a 128-unit critic, and a 64-unit prediction head.

Reported returns are on-policy training returns under the entropy-regularised policy. They are averaged over the last tenth of the budget. Figure 2 shows the learning curves of the co-learning experiments by mode. In co-learning, both agents learn, and the pair uses parameter-shared self-play. One network controls both agents, and each agent acts on its own inputs. An *arm* is one configuration of the experiment. Arms with unshared parameters appear only where stated (Sections 6.1 and 7). Three robustness experiments, registered in advance, each change one setting (Table 10). Two of them run four times the budget, one at each window. They train for 1600 iterations, with the learning-rate anneal spread over all of them. The iteration-400 point inside those runs is therefore not the end point of the 400-iteration experiments. Every reading in them is taken on the last tenth of the 1600 iterations. The third, at $w = 5$, uses unshared parameters: one actor network per agent, with gradient clipping per agent. This matches the unshared arms of the positive regime.

The scripted leader is heuristic H07 of the published game. It draws its corner uniformly at the first step of the episode and keeps it. It stands still for its private delay. It then moves toward the corner by a fixed rule: left if the corner is to its left, else up if the corner is above, else right, else down. Each action is replaced by a uniformly random one with probability 0.4. The observer's posterior scores a trajectory under this rule with this noise. A move the rule never produces is therefore scored as noise, not as evidence. For this reason the posterior is exact for the scripted leader and only a reference for a learned one.

A head on the shared recurrent network predicts the partner's next executed action by cross-entropy. Its weight is $\lambda = 0.05$ (FULL). In the fixed-partner experiment, a second weight of 0.2 is also used (FULL-HI). The head is placed on the informed agent, the uninformed agent, or both. Every arm carries the head. The no-objective arm (NOAUX) trains it at weight zero, so all arms have the same parameters by construction. Three controls keep the same head and gradient but remove the partner content:

- uniformly random labels (RANDOM);
- the agent's own next action as the label (OWN);
- the objective with its gradient stopped at the network (STOPGRAD).

Two privileged diagnostics are disclosed as such and used in training only. The first gives the partner's variable to the head as an extra input (PRIVILEGED). The second uses the variable itself as the head's target (DIRECT). Further arms are used where stated:

- the head's features fed into the policy;
- a wider recurrent network with no objective;
- the partner's variable shown to the agent as an oracle (the *shown* arm);
- the agent trained with the partner masked from its inputs (the *blind* arm).

To test whether a trained agent uses its partner, we replace the partner's inputs at evaluation with those of another episode. This input is in distribution but carries no information about the current episode. We do not use zeros, which can make an agent fail even when the inputs carry nothing (Lowe et al., 2019). The shown and blind arms serve as controls for unfamiliar inputs.

The other modelling arms are variants of published designs, not reimplementations. We state each departure below. The plain head at $\lambda = 0.05$ is a variant of the AMS design of Hernandez-Leal et al. (2019), a policy-prediction head on the shared network. That paper uses weights 0.1 to 0.5. The arms whose policy reads the head's features are variants of that paper's AMF design. They concatenate the head's hidden layer to the policy input, where AMF uses an element-wise product. We run them with and without the head's gradient stopped at the network.

The LIAM-class arm is a variant of the design stated in Papoudakis et al. (2021). A separate recurrent encoder reads the agent's own inputs. Here it is a GRU of 128 units over the agent's observation and the windowed partner action. A decoder reconstructs the partner's observation by squared error and its current action by cross-entropy. The embedding is concatenated to the policy input under a stop-gradient, so no RL gradient reaches the encoder. The encoder-decoder loss is kept apart from the RL loss. The action-only variant drops the observation head. The no-stop-gradient variant, registered in advance (Section 7), feeds the embedding to the policy without the stop-gradient. The actor-critic loss then shapes the encoder, as it shapes the head on the network in the AMF-class arm. The fresh replication (Table 11) adds two controls for the AMF-class arm: the same reading branch trained on uniformly random labels, and the same branch with the prediction loss at weight zero, so that the policy reads a branch shaped only by the RL gradient. The three reading arms have identical architecture and parameters. Our version departs from LIAM in four ways. We use recurrent PPO with a centralised critic rather than A2C, so the embedding enters the actor only. We use a single optimiser. The target is the partner's full observation, which contains its private corner. This label is privileged, and LIAM's centralised training also uses it. We do not match parameters to the no-objective arm.

The BEPAL-class target is a variant of Huang et al. (2026). The head regresses the partner's grid position at $t + 1$ to $t + 3$ by squared error. Its gradients flow into the network, and its outputs are never fed to the policy. We use weights 0.05 and that paper's 0.5. We use PPO with a centralised critic rather than their actor-critic with gated communication. We predict positions only, with no reward head and no occupancy maps. We use a three-step horizon, where the paper gives none.

The direct target uses the partner's corner as the head's label. It is available in centralised training only. Each of these arms, and the reading of each of its outcomes, was fixed before it was run.

Decodability is the held-out accuracy of a linear probe that reads the partner's variable from the agent's belief. The probe is trained on half of 256 stored episodes and tested on the other half. Legibility $L(k)$ is the held-out accuracy of a probe that predicts the informed agent's corner from its first k executed moves after its delay. This probe is trained within each run. The commit step is the first step after which the follower's nearest corner no longer changes.

Seeds are registered in advance: twenty per arm in the toy, ten in the game with a scripted or always-visible partner, and sixty in the co-learning retention experiments. No seed is dropped. Every contrast is a mean difference between two arms. It is reported with a seed-level 95% bootstrap interval and a permutation p, exact where every relabelling can be enumerated (ten seeds per arm) and Monte Carlo otherwise. Holm correction is applied within the contrast's family, declared in advance. A contrast is a *benefit* when the lower bound of its interval exceeds a practical margin. It is *equivalent* when the whole interval lies within the margin, and *inconclusive* otherwise. The margins are 0.02 of return and 0.10 of decodability or legibility. Return is the mean training return over the

last tenth of iterations. Where a run collapses, medians are read beside means. Every experiment was registered in advance with its predictions and with the reading each pattern of outcomes would receive. Every run also records a consistency check. The update pass must reproduce the rollout's action log-probabilities, and training aborts if the mismatch exceeds 10^{-4}. The largest mismatch recorded in any run of any experiment reported here is 5.7×10^{-6}.

C RESULT TABLES

Tables 2 to 14 collect the experiments that the main text quotes.

We read each null result in these tables against effects that the same instruments and margins do detect on the same seeds. With a fixed partner, hiding a relevant corner costs 0.111. With a learning partner, showing the corner gains $+0.119$, and a follower blind to its partner shows a deficit of -0.62. Swapping the partner's inputs with another episode's drops return by 0.20 to 1.25. The objective raises decodability by $+0.54$, and under a convention legibility drops by 0.40.

Two further registered experiments at $w = 8$ ask whether the leader rewards' benefit belongs to the pair or to the leader's behaviour. In the first, a follower trains against a population of frozen leaders from one arm of the replication and is evaluated with leaders it has never met (Table 13). Followers trained on either rewarded pool reach 0.707 to 0.718 with held-out leaders, 0.18 to 0.19 above followers trained on the no-objective pool, and lose at most 0.013 against their training pool. Followers of the no-objective pool lose 0.11, more than the registered 0.05, which the plan reads as partner-specific adaptation; part of the drop is the leaders', since only 13 of the 20 held-out no-objective leaders are above the threshold themselves, against all of the rewarded ones. Against frozen leaders the reading branch adds nothing ($+0.013$, $+0.012$ and $+0.002$), so its gain in the co-learning experiments needs both agents to learn. Every follower uses its leader: swapping the leader's inputs drops return to between -0.04 and 0.22.

In the second, the trained agents of the replication are paired across seeds (Table 14). Coordination is partly pair-specific: no-objective agents reach the threshold in 22 of 60 cross-seed pairings, where 45 of 60 pairs reached it in training. A progress-trained leader is easier for a no-objective follower to coordinate with ($+0.149$ $[+0.073, +0.227]$); a legibility-trained leader is inconclusive ($+0.075$ $[-0.001, +0.154]$). Followers trained with either reward do worse with an unrewarded leader (-0.142 and -0.211). These cross-play comparisons are descriptive.

In Figure 2, at 400 iterations the no-objective runs that end below the 0.5 threshold are still rising, by $+0.027$ and $+0.050$ per 100 iterations over the last 100 (18 and 36 runs), the arms with the leader reward more slowly ($+0.012$ to $+0.032$), and the reference with the leader always visible is flat ($+0.002$). The no-objective runs whose own return over iterations 361 to 400 was below the threshold (10 runs at $w = 8$, 20 at $w = 5$) reach 0.795 and 0.703 over the last tenth of four times the budget, beside 0.787 and 0.757 for the other no-objective runs and 0.751 and 0.742 for the reward arm; every slope is then under $+0.01$ per 100 iterations, and at $w = 8$ the rewarded pairs end below the unrewarded ones.

C.1 THE BINDING INTERVENTION IN FULL

The natural repair for a corner that is "retained but not acted on" is to make the policy follow its own belief. After the end of the window we pull the follower's action toward the move that leads to the corner its head decodes. That corner is held constant in the gradient. At weight 0.05 the pull changes return by -0.009, -0.007 and -0.000 across the three windows, each inside the equivalence margin. At weight 0.2 it is a deficit at every window, -0.041, -0.033 and -0.033, and the follower commits about one step *later* rather than sooner. The same pull toward a head trained on uniform labels collapses the pairs, by -0.29 to -0.17. The decoded corner therefore does carry the partner's variable. These results are consistent with the policy using that information better than a hard pull toward the head's most likely corner can.

We also ran the reference that this repair assumes. From the end of the window, the follower is shown the corner that its own exact posterior makes most likely, and this corner is then frozen. The follower is thus handed the best single declaration its information supports. This arm earns *less* than the untouched follower at every window, by 0.033, 0.047 and 0.024. It acts as if the declared corner

Table 2: Known-answer suite: scripted followers with stated recurrent states B against the scripted leader under the target-only rule, 2,048 episodes per case (Monte Carlo half-width 0.022 on the accuracies; 0.043 on their differences and on return). C is the compatibility of the follower's declaration at step τ; $V(H_\tau)$ and $V(B)$ the best compatibility attainable from the history and from the state; T_1, T_2, T_3 the unavailable, representation and unused terms; $I(U;Z)$ the mutual information between the declared and the private corner, in units of $\log 4$. Case (a) uses the permissive rule and a leader that walks toward the follower; its C is written from the script. $V(B)$ is set equal to $V(H)$ by the stated state in cases (b) to (d) and measured by a held-out decoder in case (f).

case	follower at τ	C	$V(H_\tau)$	$V(B)$	T_1 / T_2 / T_3	$I(U;Z)$	return	success
(a) perfect convention	fixed corner, permissive rule	1.00	–	–	0 / 0 / 0	0.00	$+1.00$	1.00
(b) conditional response	declares $\arg\max_z p_\tau$ at $\tau = 8$	0.81	0.81	0.81	0.19 / 0.00 / 0.00	0.59	$+0.44$	0.81
(b′) free timing	declares when $\max_z p_\tau \geq 0.8$	0.89	0.89	0.89	0.11 / 0.00 / 0.00	0.71	$+0.61$	0.89
(b″) as (b′), revisable	re-declares every step	0.89	0.89	0.89	0.11 / 0.00 / 0.00	0.71	$+0.79$	1.00
(c) correlated, incompatible	corner opposite the argmax	0.03	0.81	0.81	0.19 / 0.00 / 0.79	0.59	-0.95	0.03
(d) retained, ignored	holds p_τ, declares corner 0	0.25	0.81	0.81	0.19 / 0.00 / 0.56	0.00	-0.56	0.25
(d′) unavailable to B	holds nothing, declares corner 0	0.25	0.81	0.25	0.19 / 0.56 / 0.00	0.00	-0.56	0.25
(e) compatible, poor execution	corner shown, executes with noise 0.4	1.00	1.00	1.00	0 / 0 / 0	1.00	$+0.85$	1.00
(e_0) oracle	corner shown, shortest path	1.00	1.00	1.00	0 / 0 / 0	1.00	$+0.94$	1.00
(f) reactive tracking	corner nearest the leader's cell	0.70	0.81	0.74	0.19 / 0.07 / 0.05	0.40	$+0.24$	0.70

Table 3: The commitment task against its closed forms: eight settings of (g, p), six arms, 20 seeds per arm, mean return over the last tenth of training. V_X, V_H and V_Z are the exact values with reactive, history and type information; R, H, M, U, M-direct and P should reach V_X, V_H, V_H, V_H, V_H and V_Z. The last column is the history learner's gain over the reactive learner with its seed-level 95% interval, beside the history value $V_H - V_X$. At $g = 1$, $p = 1$ one H seed and one P seed collapsed to 0.5 (medians 1.000); a follow-up with 100 fresh seeds per arm had 0 collapses (Appendix A).

cell (g, p)	$V_X / V_H / V_Z$	$V_H - V_X$	R	H	M	U	M-direct	P	H − R [95% CI]
0, 1	1 / 1 / 1	0	0.9998	0.9998	0.9998	0.9999	0.9998	0.9999	$+0.0000\ [-0.0001, +0.0002]$
0, 0.5	1 / 1 / 1	0	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	$-0.0000\ [-0.0001, +0.0000]$
0.5, 1	0.75 / 0.75 / 0.75	0	0.7494	0.7493	0.7493	0.7494	0.7493	0.7494	$-0.0001\ [-0.0022, +0.0021]$
0.75, 0.65	0.625 / 0.625 / 0.875	0	0.6240	0.6236	0.6240	0.6238	0.6227	0.8747	$-0.0003\ [-0.0036, +0.0029]$
1, 0.5	0.5 / 0.5 / 1	0	0.5009	0.5004	0.5000	0.4997	0.5003	0.9998	$-0.0005\ [-0.0050, +0.0040]$
0.75, 0.8	0.625 / 0.725 / 0.875	0.1	0.6242	0.7255	0.7260	0.7255	0.7260	0.8747	$+0.1014\ [+0.0987, +0.1040]$
1, 0.8	0.5 / 0.8 / 1	0.3	0.4988	0.8009	0.8013	0.8013	0.8014	0.9998	$+0.3022\ [+0.2981, +0.3057]$
1, 1	0.5 / 1 / 1	0.5	0.4994	0.9752	0.9998	0.9998	0.9997	0.9753	$+0.4758\ [+0.4249, +0.5025]$

Table 4: When only the leader's corner pays and both agents learn, with memory trained through time (seeds 30 to 39, 10 per arm). The centralised critic receives the leader's corner in every arm; the follower's policy never does. Levels are the mean return over the last tenth of training in a repetition on a second hardware platform; the original run gives the same levels within 0.003. Differences are against the no-objective pair of the same architecture, with seed-level 95% intervals (margin 0.02); every objective difference is equivalent. Swap: return when the leader's inputs are replaced at evaluation by another episode's. The blind arm varies widely across seeds (standard deviation 0.18 to 0.22), so it is a floor rather than a level.

arm	GRU: return	GRU: contrast vs no objective	attention: return	attention: contrast
co-learning, no objective	0.821	ref.	0.815	ref.
objective on the leader	0.82	-0.001 $[-0.003, +0.001]$	0.82	$+0.002$ $[-0.001, +0.004]$
objective on the follower	0.82	$+0.001$ $[-0.001, +0.002]$	0.82	$+0.002$ $[+0.000, +0.005]$
objective on both	0.82	-0.001 $[-0.003, +0.002]$	0.82	$+0.002$ $[-0.000, +0.004]$
random labels on the leader (control)	0.82	-0.000 $[-0.002, +0.002]$	0.81	-0.002 $[-0.010, +0.003]$
unshared parameters, no objective	0.82	$+0.003$ $[+0.001, +0.004]$	0.82	–
unshared, objective on the leader (vs unshared no objective)	0.82	-0.000 $[-0.002, +0.001]$	0.82	$+0.002$ $[-0.000, +0.004]$
scripted informed leader, learned follower	0.821	-0.000 $[-0.001, +0.001]$	0.818	$+0.003$ $[+0.001, +0.005]$
corner shown to the follower (oracle)	0.940	$+0.119$ $[+0.117, +0.120]$	0.94	$+0.124$ $[+0.122, +0.126]$
follower trained without leader inputs (blind)	0.21	-0.616 $[-0.746, -0.488]$	0.16	-0.655 $[-0.753, -0.555]$
every hidden-corner arm under a swapped leader	-0.32 to -0.41		-0.34 to -0.39	

Table 5: Fixed scripted partner with memory trained through time (hard regime: execution noise 0.4, private delay up to 6 steps, waiting has a cost; seeds 40 to 49, 10 per arm). Return is the mean over the last tenth of training. Decodability is the held-out accuracy of a linear probe reading the leader's corner from the follower's memory at steps 1 and 3 (chance 0.25); the deficit is the memory's gap to the exact posterior at the commit step. All twenty return contrasts against no objective are within 0.003. When the corner is shown but irrelevant, the objective makes it decodable although the reward never uses it; when it is hidden and relevant, no arm changes what the memory holds. The registered probe step, step 5, is undefined in most shown-corner episodes, which end after four steps, so steps 1 and 3 are reported instead.

cell	objective	return	decodability, step 1	step 3	deficit at commit	commit step
corner shown, irrelevant	no objective	0.8148	0.42	0.46	0.56	2.4
	next action, $\lambda = 0.05$	0.8148	0.96	0.90	0.20	2.4
	next action, $\lambda = 0.2$	0.8148	0.98	0.96	0.12	2.4
	stop-gradient (control)	0.8148	0.42	0.48	0.56	2.4
	random labels (control)	0.8148	0.47	0.48	0.57	2.4
corner hidden, irrelevant	no objective	0.8148	0.24	0.24	0.32	2.4
	next action, $\lambda = 0.05$	0.8148	0.24	0.30	0.32	2.4
	next action, $\lambda = 0.2$	0.8148	0.25	0.31	0.28	2.4
	stop-gradient, random labels, privileged head	0.8148	0.23 to 0.26	0.25 to 0.29	0.31 to 0.33	2.4
	direct target: the corner	0.8148	0.26	0.38	0.26	2.4
corner hidden, relevant	no objective	0.7034	0.28	0.37	0.14	8.4
	next action, $\lambda = 0.05$ and 0.2	0.7036, 0.7042	0.27 to 0.28	0.37 to 0.40	0.13 to 0.14	8.4
	stop-gradient, random labels, privileged head	0.7032 to 0.7034	0.27 to 0.28	0.37 to 0.40	0.13 to 0.14	8.4 to 8.5
	direct target: the corner	0.7048	0.28	0.39	0.12	8.4
corner shown, relevant	no objective	0.8146	1.00	1.00	0.03	2.5
	every objective and control	0.8146 to 0.8147	1.00	0.99 to 1.00	0.02 to 0.04	2.4 to 2.5

Table 6: Retention with the scripted leader hidden after the follower's first w moves (target-only rule, GRU followers, seeds 500 to 509, 10 per arm). No-objective rows give levels: return, decodability of the corner from memory at the end of the window and ten steps later, and the commit step relative to the end of the window. The computed value of one declaration at the end of the window is 0.63, 0.79 and 0.91, and the best accuracy measured on the same episodes is 0.63, 0.77 and 0.86. Other rows give differences from no objective with seed-level 95% intervals (margins 0.02 of return and 0.10 of decodability). Every return and end-of-window decodability difference is equivalent; decodability ten steps later and the commit step are inconclusive at ten seeds, with intervals up to ± 0.4 steps on the commit step. The head is at $\lambda = 0.05$. The swap lowers the no-objective follower's return by 0.20, 0.34 and 0.60 and the blind arms' by 0.00 to 0.03. References: leader visible throughout 0.8215 (decodability 0.74 at the same alignment, commit at $+1.5$); corner shown 0.940; blind 0.58.

window	arm	return	decodable at window end	decodable at $+10$	commit relative to window end
$w = 3$	no objective (level)	0.708	0.58	0.48	$+8.5$
	next action	-0.002 $[-0.007, +0.003]$	-0.007 $[-0.038, +0.024]$	$+0.015$	$+0.12$ $[-0.33, +0.55]$
	direct target: the corner	$+0.004$ $[-0.000, +0.009]$	$+0.011$ $[-0.023, +0.042]$	$+0.007$	-0.19
	head features into the policy	$+0.004$	$+0.002$	$+0.021$	-0.14
	wide trunk, no objective	$+0.004$	$+0.005$	-0.012	$+0.02$
	random labels; own action	$+0.003$; $+0.001$	$+0.006$; $+0.003$	-0.003; -0.033	-0.09; -0.05
$w = 5$	no objective (level)	0.752	0.74	0.45	$+4.4$
	next action	$+0.003$ $[+0.001, +0.005]$	-0.001 $[-0.033, +0.030]$	-0.012	-0.12 $[-0.44, +0.21]$
	direct target: the corner	$+0.006$ $[+0.003, +0.008]$	-0.004 $[-0.030, +0.023]$	-0.023	-0.20
	head features into the policy	$+0.005$	-0.009	-0.031	-0.14
	wide trunk, no objective	$+0.005$	$+0.001$	-0.051	-0.15
	random labels; own action	$+0.003$; -0.001	-0.005; -0.011	-0.006; $+0.014$	$+0.02$; -0.03
	attention: no objective (level)	0.741	0.69	0.45	$+4.9$
	attention: next action	-0.004 $[-0.012, +0.003]$	$+0.009$ $[-0.023, +0.040]$		
$w = 8$	no objective (level)	0.781	0.79	0.48	$+0.1$
	next action	-0.004 $[-0.010, +0.002]$	$+0.014$ $[-0.025, +0.052]$	-0.096	$+0.12$
	direct target: the corner	$+0.007$ $[+0.003, +0.010]$	$+0.035$ $[+0.003, +0.066]$	$+0.045$	-0.12
	head features into the policy	$+0.002$	$+0.005$	$+0.031$	-0.10
	wide trunk, no objective	$+0.001$	$+0.018$	-0.027	$+0.07$
	random labels; own action	$+0.001$; $+0.001$	-0.001; $+0.020$	-0.058; $+0.029$	$+0.11$; -0.21

Table 7: Two controls at $w = 5$ with both agents learning (seeds 800 to 859, 60 per arm, 420 runs), each against this experiment's own no-objective arm. That arm reaches the threshold in 32 of 60 runs, against 24 on the same seeds in the replication; the code is identical and the difference comes from floating-point arithmetic on different hardware (Appendix H.1). Top: the number of runs with task return above 0.5, the rate difference with its Newcombe interval, and the return difference with its seed-level 95% interval (margin 0.02). Bottom: the leader's legibility and the follower's decodability and commit step at the end of the window (margin 0.10). A reward for distance covered toward the goal, with no observer involved, lifts as many runs above the threshold as the legibility reward and raises legibility and decodability about as much, so the two rewards cannot be separated in this game. Of the four variants of the follower's head, only the two whose features are fed to the policy change return; the variant named decisive in advance (gradient stopped, features read) misses its 0.2 rate margin. The last rows of each block repeat the legibility arms of the replication on the same seeds, against that experiment's own no-objective arm.

arm	trunk	policy reads	above 0.5 (rate minus ref)	return
no objective (level)			32 of 60	0.467
next action, plain	shared	no	27 (-0.08 $[-0.25, +0.09]$)	-0.021 $[-0.092, +0.050]$
next action, dead end	stopped	no	27 (-0.08 $[-0.25, +0.09]$)	-0.057 $[-0.136, +0.021]$
next action, features into the policy	shared	yes	46 ($+0.23$ $[+0.06, +0.39]$)	$+0.122$ $[+0.058, +0.188]$
next action, decoupled (named decisive)	stopped	yes	42 ($+0.17$ $[-0.01, +0.33]$)	$+0.100$ $[+0.032, +0.168]$
progress reward, $\beta = 0.5$			57 ($+0.42$ $[+0.27, +0.55]$)	$+0.165$ $[+0.104, +0.224]$
progress reward, $\beta = 2.0$			60 ($+0.47$ $[+0.33, +0.59]$)	$+0.148$ $[+0.098, +0.201]$
replication: no objective (level)			24 of 60	0.415
legibility reward, $\beta = 0.5$			58 ($+0.57$ $[+0.42, +0.68]$)	$+0.260$ $[+0.203, +0.317]$
legibility reward, $\beta = 2.0$			60 ($+0.60$ $[+0.46, +0.71]$)	$+0.256$ $[+0.202, +0.311]$

arm	probe $L(3)$	decodable at window end	commit relative to window end
next action, plain	-0.002 $[-0.026, +0.021]$	$+0.019$ $[-0.010, +0.047]$	$+0.11$ $[-0.85, +1.04]$
next action, dead end	-0.010 $[-0.034, +0.014]$	-0.006 $[-0.034, +0.022]$	$+0.52$ $[-0.53, +1.59]$
next action, features into the policy	$+0.040$ $[+0.017, +0.064]$	$+0.080$ $[+0.052, +0.108]$	-2.15 $[-3.03, -1.28]$
next action, decoupled (named decisive)	$+0.021$ $[-0.004, +0.046]$	$+0.065$ $[+0.035, +0.094]$	-1.53 $[-2.42, -0.62]$
progress reward, $\beta = 0.5$	$+0.075$ $[+0.055, +0.095]$	$+0.143$ $[+0.121, +0.165]$	-2.78 $[-3.58, -2.00]$
progress reward, $\beta = 2.0$	$+0.074$ $[+0.054, +0.092]$	$+0.165$ $[+0.143, +0.186]$	-1.50 $[-2.22, -0.83]$
legibility reward, $\beta = 0.5$	$+0.093$ $[+0.073, +0.113]$	$+0.157$ $[+0.134, +0.179]$	-3.84 $[-4.65, -3.05]$
legibility reward, $\beta = 2.0$	$+0.097$ $[+0.077, +0.116]$	$+0.172$ $[+0.148, +0.196]$	-3.33 $[-4.06, -2.60]$

were certain, and so it pays in the roughly 37%, 21% and 9% of episodes where the declaration ceiling predicts the declaration to be wrong. The untouched follower instead keeps a calibrated belief and hedges. The declaration ceiling of Section 3 therefore bounds a single irrevocable declaration, not continued play. Because the arm discards the posterior's confidence, it is not a ceiling on how well the information can be used. We did not test whether a follower given its full posterior, or the exact solution of its decision problem, would exceed the untouched follower.

D PROPOSITIONS, PROOFS AND THE PRE-REGISTERED READINGS

Throughout, Z and the action set are finite, and H is fixed. We write $q(z) = P(Z = z \mid H)$ for the posterior and $Q(z, u)$ for the action value.

Proposition 1 (zero benefit) $\mathrm{VPI}(H) = 0$ *if and only if some* $u^\star$ *lies in* $\bigcap_{z:q(z)>0} \arg\max_u Q(z, u)$.

Proof. If such $u^\star$ exists, $\sum_z q(z) \max_u Q(z, u) = \sum_z q(z) Q(z, u^\star) \le \max_u \sum_z q(z) Q(z, u) \le \sum_z q(z) \max_u Q(z, u)$, so all are equal. Conversely, let u_0 attain $\max_u \sum_z q(z) Q(z, u)$. Then $\sum_z q(z) \left[\max_u Q(z, u) - Q(z, u_0)\right] = \mathrm{VPI}(H) = 0$, and every summand is non-negative. So $Q(z, u_0) = \max_u Q(z, u)$ for every z with $q(z) > 0$. □

Now take $Q = K$, the compatibility predicate. Then $\arg\max_u K(z, \cdot)$ is the set of corners compatible with z. The value is therefore zero if and only if the compatible sets of all supported corners intersect. The permissive rule satisfies this for every history. The target-only rule violates it whenever the posterior supports two corners.

Table 8: The $w = 8$ co-learning experiment in full (parameter-shared self-play, 400 iterations, seeds 800 to 859, 60 per arm, 660 runs), against its own no-objective arm. Counts are runs whose task return over the last tenth exceeds 0.5, with the rate difference and its Newcombe interval. Return differences carry seed-level 95% intervals and a verdict at the 0.02 margin (B benefit, I inconclusive, D deficit), with Holm correction across the ten arms. The last column is the difference of means among runs above the threshold, descriptive and without an interval. The reward brings every run above the threshold at both weights. At $\beta = 0.5$ those runs end where the no-objective runs above the threshold end; at $\beta = 2.0$ they end lower (0.647 to 0.691, below 26 of the 42 no-objective runs above the threshold). The three placements of the head and the two content-free controls lie within 8 of 60 of the no-objective count, the spread between two repetitions of the no-objective arm at $w = 5$ (Appendix H.2). The head is at $\lambda = 0.05$ and the controls at $\beta = 2.0$. The no-objective and reward counts are the same for any threshold from 0.5 to 0.6.

arm	above 0.5 (rate minus ref)	return (verdict)	within mode
no objective (level)	42 of 60	0.591	0.708
head on the follower	36 (-0.10 $[-0.26, +0.07]$)	-0.026 $[-0.091, +0.042]$ I	$+0.005$
head on the leader	44 ($+0.03$ $[-0.13, +0.19]$)	$+0.006$ $[-0.062, +0.074]$ I	-0.004
head on the leader, random labels	40 (-0.03 $[-0.20, +0.13]$)	-0.014 $[-0.086, +0.056]$ I	$+0.006$
legibility reward, $\beta = 0.5$	60 ($+0.30$ $[+0.18, +0.43]$)	$+0.131$ $[+0.085, +0.181]$ B	$+0.013$
legibility reward, $\beta = 2.0$	60 ($+0.30$ $[+0.18, +0.43]$)	$+0.083$ $[+0.038, +0.133]$ B	-0.034
reward from another episode	44 ($+0.03$ $[-0.13, +0.19]$)	-0.092 $[-0.151, -0.031]$ D	-0.128
reward for an unused goal	47 ($+0.08$ $[-0.07, +0.24]$)	-0.033 $[-0.094, +0.030]$ I	-0.071
leader visible throughout	60 ($+0.30$ $[+0.18, +0.43]$)	$+0.230$ $[+0.185, +0.280]$ B	$+0.113$
blind follower, with the head	3 (-0.65 $[-0.76, -0.50]$)	-0.311 $[-0.372, -0.248]$ D	
blind follower, no objective	3 (-0.65 $[-0.76, -0.50]$)	-0.392 $[-0.460, -0.322]$ D	

Proposition 2 (the attainable gain) *Let $u_0 \in \arg\max_u \sum_z q(z)Q(z,u)$ be a best action under the posterior alone, and let $E = \{z : q(z) > 0,\ u_0 \notin \arg\max_u Q(z,u)\}$ be the partner variables for which u_0 is not optimal. Then $\mathrm{VPI}(H) = \sum_z q(z)\,\Delta(z) = q(E)\,\mathbb{E}[\Delta(Z) \mid Z \in E]$ with $\Delta(z) = \max_u Q(z,u) - Q(z,u_0)$, and $\Delta(z) = 0$ for every $z \notin E$.*

Proof. Immediate from (1). □

With $Q = K$, $\Delta \in \{0,1\}$, so $\mathrm{VPI}_K(H) = q(E) = 1 - \max_z q(z)$ under the target-only rule. This is why the value can be computed from the posterior.

Proposition 3 (predictable is not useful) *Let $Z = (Z_r, Z_i)$ with $Q(z,u) = Q(z_r,u)$, and let H' be evidence with $I(Z_r; H' \mid H) = 0$, equivalently $P(Z_r \mid H, H') = P(Z_r \mid H)$ almost surely. Then $\mathrm{VPI}(H, H') = \mathrm{VPI}(H)$, while $I(Z; H' \mid H) = I(Z_i; H' \mid Z_r, H)$ can be any value up to $H(Z_i \mid Z_r, H)$, and over a family of tasks with growing $|\mathcal{Z}_i|$ can be made arbitrarily large.*

Proof. Substituting $Q(z_r,u)$ into (1) makes both terms functions of $P(Z_r \mid H, H')$. The hypothesis leaves this equal to $P(Z_r \mid H)$. The information statement follows from the chain rule $I(Z; H' \mid H) = I(Z_r; H' \mid H) + I(Z_i; H' \mid Z_r, H)$. Its first term is zero by hypothesis. Its second term is bounded by $H(Z_i \mid Z_r, H)$ on a finite alphabet. □

We state the hypothesis on Z_r directly because a weaker-looking one does not suffice. That weaker condition is conditional independence of Z_r and H' given H and Z_i. For a counterexample, take Z_r uniform on $\{0,1\}$, $Z_i = Z_r$ and $H' = Z_i$. This satisfies the weaker condition, yet H' determines Z_r. What must be ruled out is any indirect Bayesian shift in $P(Z_r \mid \cdot)$. Without that, evidence about Z_i moves the posterior over Z_r, and the value moves with it. In the toy, the settings with $g = 0$ realise the proposition with Z_r empty. In Section 6.2, the condition in which the corner is shown but irrelevant to the follower's reward realises it in a trained agent.

D.1 Where the decision value is lost, and the known-answer suite

Table 2 evaluates scripted followers with stated recurrent states against the scripted leader. Every term is therefore known by construction, and the measures are checked before they are applied to a learned agent. Cases (b) and (c) have the same mutual information between the declared and the private corner, but compatibility 0.81 and 0.03. Cases (d) and (d′) have the same compatibility and

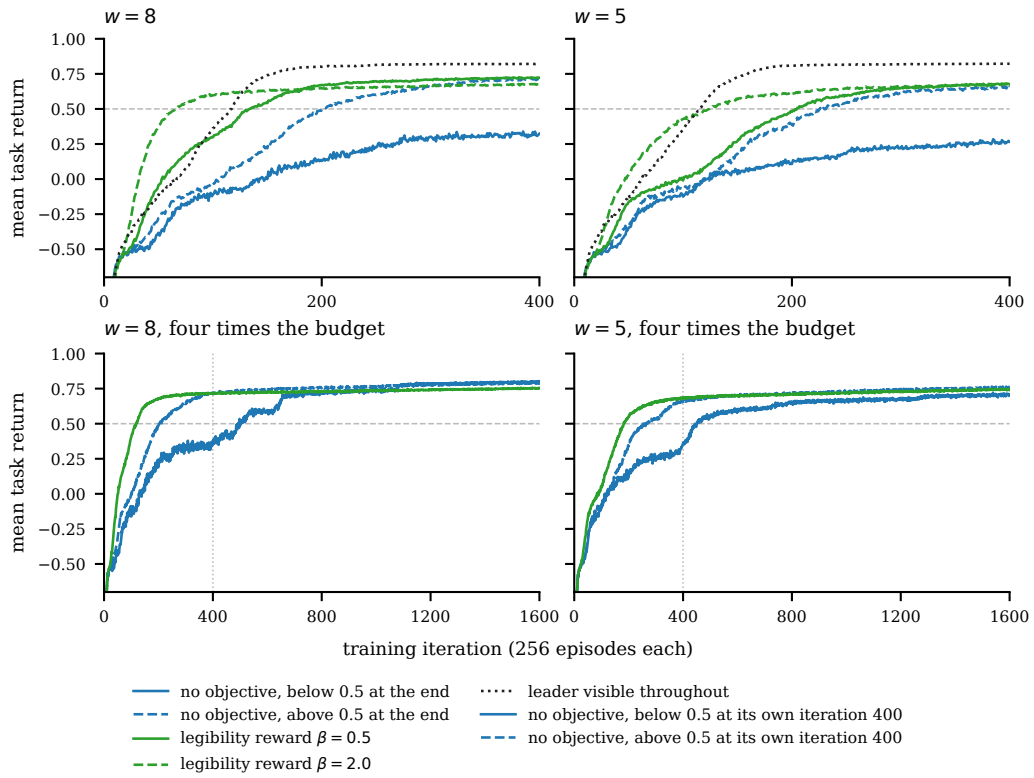

Figure 2: Training curves of the co-learning experiments by outcome, at $w = 8$ (left) and $w = 5$ (right), for 400 iterations (top) and four times the budget (bottom; dotted line at iteration 400). Given four times as long, the no-objective runs that were below the 0.5 threshold at iteration 400 reach the level of the others, and every curve flattens: the low runs are slow, not stuck (Table 10).

return. They differ only in whether the information reached the state and was ignored, or never reached the state. Two more differ only in execution. Return and mutual information each misorder a pair that the value orders. Mutual information ranks (c) above (d), 0.59 against 0.00, whereas compatibility ranks (d) above (c), 0.25 against 0.03. Return ranks (f) above (d), $+0.24$ against -0.56, whereas the value in the state ranks (d) above (f), $V(B) = 0.81$ against 0.74. Return penalises (d) for never using what it holds, and credits (f) for arriving anyway.

D.2 The retention readings, fixed before the run

The table below gives, for each pattern of outcomes, the reading we fixed before the retention run. In it, "the close" is the end of the window during which the follower sees the leader.

decodability at the close	commitment given the belief	return	reading
below the ceiling with no objective; raised toward it by the next-action and direct-target arms; controls flat	follows decodability	rises with it	retention lost; the objective addresses it through its target
below with no objective; raised by the direct target only	follows the direct target	rises only there	retention lost; the next-action target does not train the missing quantity
below with no objective; raised by every auxiliary arm, own-action and random labels included	follows decodability	rises everywhere	a generic auxiliary-training effect, not teammate modelling
at the ceiling in every arm	below the attained decodability	flat	retained but unused; no representation objective can fix it
at the ceiling and used; return at the leader-visible level	at the ceiling	flat	no shortfall at this window

The table is reproduced exactly as it was fixed before the run. It records what was pre-registered, not what the paper concludes. In particular, the causal reading in its fourth row is tested in Section 7, and the results there do not support it. A positive result for teammate modelling was to be claimed only on the first row. A null result against the objective was to be claimed only on the second and fourth rows. The experiment landed on the fourth row. Its commitment column was measured by the commit timing, because the accuracy measure is degenerate under the target-only rule (Section 6.3).

D.3 The three-term identity and its bounds

Consider a learner whose action depends on the history only through its recurrent state B. Then $V(Z) \geq V(H) \geq V(B) \geq C$. Here $V(\cdot)$ is the best compatibility any decoder of that information attains, and C is the accuracy of the decision the learner makes. The shortfall $1 - C$ splits into three non-negative terms: unavailable, $1 - V(H)$; lost in the representation, $V(H) - V(B)$; and retained but unused, $V(B) - C$. The first term is computed with a scripted partner. The other two are only partially identified. A linear probe on B gives a lower bound p on $V(B)$, and the learner's own decision gives another, c. So $V(H) - \max(p, c)$ bounds the representation loss from above, and $\max(0, p - c)$ bounds the unused term from below. Two learners with identical return can differ only in which term they lose to, and no return measurement separates them.

E The permissive-rule experiment in full

Under the permissive rule any shared corner pays. When both agents learn under this rule, all 140 co-learning pairs settle on one meeting corner. The follower commits to that corner in every episode. The corner agrees with the leader's private corner only at chance. The leader's route carries no information about it: $L(3)$ is 0.25, against 0.65 for the scripted leader under identical noise. The follower commits at step 2.5, before the leader's delay has ended. With the scripted leader it commits at step 8.4. In these runs the seed determines which corner a pair adopts. Return is set by the corner adopted: every run ends at 0.83 or at 0.58, with nothing between. No placement of the objective changed the convention, the legibility, the commit step or the return. The setting with the corner shown gives the same result. There the follower decodes the corner at 0.79 to 0.85, and the pair still meets at the convention corner. The value of the partner's information is zero here because of the partner each pair reached, not because of the task. Against the scripted leader the same rule gives that information a value of $3/4$. One probe does move. With the objective on the leader, the follower's state becomes more decodable from the leader's memory, by $+0.170$ $[+0.131, +0.209]$ with unshared parameters, while no behaviour changes. This is the fixed-partner dissociation of Section 6.2 again, seen from the leader's side.

F OVERCOOKEDV2 WITH A FROZEN PARTNER

We test whether the gridworld results hold in a published cooperative benchmark with a private variable.

OvercookedV2 (Gessler et al., 2025) (layout *asymmetric advantages, recipes right*) is a two-room kitchen in which the dish to deliver, the recipe, is drawn at random and redrawn after each correct delivery; a wrong delivery costs 20. The partner is a policy trained beforehand with a full view of the kitchen and then frozen, so it always knows the recipe. The learner sees only the cells within two steps of itself, so the recipe reaches it mainly through the partner's cooking. Training uses 391 iterations of 64 episodes of 400 steps, with a shaping reward that fades to zero halfway; every learner's memory is trained through time, as in the main text.

Twenty fresh seeds per arm: no objective; the next-action head on the learner; the same head on random labels; the same head with its features read by the policy; and the head predicting the recipe itself from a privileged training label. The primary outcome, fixed before the run, is the sparse return over the last tenth of training, compared with no objective at a margin of 5, a quarter of the cost of hiding the recipe measured when the setting was built.

No arm changes return by the margin (Table 15). The random-label head and the recipe target are equivalent to no objective, with their whole intervals inside the margin; the plain head and the feature-reading head are inconclusive, with negative point estimates and no gain above 1.4. The learner settles into one of two roles: in most runs it also loads pots (at least 2.4 loads per episode, final return about 206), and in the rest it only delivers (at most 0.06 loads, about 192). Both next-action heads leave fewer runs in the loading role, 13 of 20 against 18 of 20 with no objective (Fisher $p = 0.13$), which accounts for their lower means; the random-label head and the recipe target do not (19 and 18 of 20). No run of any arm reaches a sparse return of 215.

The representation does move. A linear probe on the learner's memory reads the recipe, averaged over the episode, at 0.38 with no objective. The next-action head raises this to 0.51 ($+0.130$ $[+0.113, +0.146]$, clearing the registered 0.10 margin) and the feature-reading head to 0.50, while random labels leave it at 0.38 (-0.001 $[-0.021, +0.019]$). Predicting the recipe itself raises it to 0.81 ($+0.432$ $[+0.416, +0.446]$), with return equivalent to no objective. The registered dissociation reading is met: the objective changes what the memory holds, and return does not follow.

In this benchmark no prediction objective improved coordination, and neither did representing the recipe itself. The extra branch read by the policy, which helped in the gridworld (Table 11), did not help here. The gridworld's finding that prediction adds nothing to return transfers; the finding that an extra actor-readable branch speeds learning does not, and we state it for the gridworld only.

A companion experiment, also planned before it ran, repeats three arms with the whole kitchen in view (same partner, budget and seeds). Seeing the recipe raises the no-objective return by $+11.1$ $[+5.0, +16.8]$, so, with memory trained through time, hiding the recipe costs about twice the margin used above. We expected the objectives to be equivalent here; neither is. The next-action head is inconclusive ($+3.35$ $[-4.06, +10.44]$), and the feature-reading head is inconclusive with a positive interval ($+7.53$ $[+2.09, +13.34]$, Holm $p = 0.031$): it removes most of the spread across seeds (223.1 ± 2.8 against 215.6 ± 13.0). This cell has no random-label or loss-free branch, so we do not attribute that gain to predicting the partner.

G A PROTOCOL FOR EVALUATING PARTNER-MODELLING CLAIMS

Each step below is motivated by a result in this paper in which omitting the step would have led to a wrong conclusion.

1. Ask whether the partner's variable can change a decision. Compute or bound its decision value for the task and the partner's policy (Section 3). Where the value is zero, a prediction objective can change what the agent's memory contains but cannot improve its best decision (Section 6.2).
2. Measure the baseline before adding an objective. Recurrent agents without any objective already reached the exact value in the commitment task and the level of a scripted informed

partner in the gridworld (Sections 5 and 6.1); an objective can only add what the baseline leaves.

3. Match every change with a content-free control. Controls with the same parameters and gradient but no partner content separate the effect of partner information from that of an extra loss. Any new input to the policy needs its own control: the head whose features the policy reads helped, but the same branch trained on random labels, or with no prediction loss, helped as much (Table 11).
4. Test use, not only presence. Decodability shows what memory contains, not what the policy uses; replacing the partner's inputs with another episode's tests use without changing their distribution (Section 4).
5. Separate speed from final level. A difference at one training budget can disappear with more training: the leader reward's advantage fell from $+0.131$ to -0.038 at four times the budget (Table 10), and the feature-reading head's from $+0.167$ to $+0.005$ (Section 7).
6. Report the full distribution of runs. Where outcomes are bimodal, a mean hides whether an intervention moves runs between modes or within one; counting runs above a threshold fixed in advance shows which (Section 7).
7. Fix the readings before the run. Two of our conclusions are negative for the tempting interpretation: the leader reward is not signalling (Section 7) and the feature-reading head does not gain from partner content (Table 11). Both were decided by rules written before the data existed (Appendix H.2).

H Reproducibility and changes to the analysis plans

Every experiment was specified before it ran: its arms, seeds, margins and how each pattern of outcomes would be read. Every number in the paper is computed by one statistics module from per-run summaries released with the code. Training is not bit-for-bit reproducible across hardware, so our standard for a reproduction is agreement of verdicts, and of arm means within the original between-seed spread.

H.1 Replication

- The commitment task's core, boundary and collapse experiments (480, 480 and 500 runs) were re-run with the same arms and seeds. Arm means agree within 0.0011 in 21 of 24 core arms, and within 0.0005 in every boundary arm, with every verdict unchanged (history gains $+0.1015$ and $+0.3028$ against $+0.1014$ and $+0.3022$). The three exceptions all involve collapsed seeds, and every arm agrees on the median. Collapse from the optimum is rare (two of forty no-objective runs in the core experiment, one in 500 in each run of the follow-up) and no arm differs in its rate, so we do not claim that any objective stabilises training.
- The positive regime was re-run on a second hardware and software platform: the informed levels agree within 0.003 (0.8212, 0.8154, 0.8212, 0.8180, 0.9397 against 0.821, 0.817, 0.821, 0.818, 0.940; Table 4 and the text report the first list), every objective contrast is equivalent, and every hidden-corner arm falls below zero under the swap on both platforms. The partner-blind learner varies widely (0.21 ± 0.22 and 0.16 ± 0.18 against 0.248 and -0.021), so we quote it only as a floor.
- The fixed-partner, permissive-rule and retention experiments were re-run with the same arms and seeds (240, 170 and 350 runs): every arm lies within 0.0064 of its original and every verdict is unchanged.
- The central ordering of Section 7 was replicated on fresh seeds in one experiment (Table 11). Repeated no-objective arms at the same window vary by about 8 of 60 runs above the threshold (35 to 45 of 60 at $w = 8$, 24 to 32 at $w = 5$), which is why every comparison in Section 7 is made against the no-objective arm of its own experiment.

H.2 Changes to the analysis plans

Each change below is reported against the plan that was fixed before the run.

The first co-learning experiment at $w = 8$ named a collapse statistic as its reliability outcome: a run whose peak return is above 0 and whose final return is at most 0.5. Because it conditions on the peak, it misses runs whose return never rose (1 of 18 for no objective, 1 of 16 and 2 of 20 for two head placements, and 3 and 11 of 57 for the two blind arms; none for the reward arms). We count runs above 0.5 directly instead; the threshold and the 0.2 margin on the rate are the planned ones, and the direct count was fixed before the $w = 5$ experiment ran.

The plan of the first $w = 8$ experiment claimed a positive result for the reward only if the legibility probe also cleared its 0.10 margin. The rate cleared its margin and the probe did not ($+0.045$ to $+0.081$ across probe lengths). The $w = 5$ plan, written after that outcome and before its run, fixed the reading used here: a reliability gain with the probe inside its margin is a gain by a route the probe does not resolve. Its progress control then ruled out the signalling interpretation (Section 7).

The plan of the controls experiment required the no-objective arm to lie within 0.05 of the earlier $w = 5$ experiment on the same seeds (0.4152, 24 of 60). It reached 0.4670 (32 of 60), missing by 0.002. The training code differs only in a branch no no-objective run executes; the 16 seeds that ran on the same hardware reproduce the earlier runs exactly, and the other 44 diverge, as floating-point differences amplified by a bimodal outcome would. Every reading of that experiment is therefore made against its own no-objective arm.

The two experiments at four times the budget met their planned reading that runs below the threshold are slow rather than stuck (57 and 60 of 60 no-objective runs above the threshold at 1600 iterations). The unshared-parameter experiment met its reading that the interventions do not depend on parameter sharing, and also its reading for an unshared baseline outside the shared range: its no-objective arm, 40 of 60, lies $+0.21$ $[+0.07, +0.33]$ above the pooled shared runs, and we report both (Table 10). The LIAM-style arm without its stop-gradient fell between two readings: its return cleared the margin, and its rate gain, $+0.18$ $[+0.02, +0.34]$, missed the 0.2 margin although the Fisher test passed ($p = 0.047$). In the replication, the planned rule attributes the gain of the feature-reading head to the extra branch rather than to partner content, because a random-label branch and a loss-free branch match it (Table 11). The capacity control met its planned reading that the gain comes from the readout and not from added parameters (Table 12). The feature-reading head at four times the budget met its planned speed reading: equivalent to its own no-objective pair at 1600 iterations ($+0.005$ $[+0.000, +0.010]$), after a benefit at 400.

I What the agents learn: environments, curves, distributions and episodes

This appendix shows the material behind the tables. It draws the two environments as the learners see them. It gives the training curves of the fixed-partner experiments and of the 400-iteration parameter-shared co-learning experiments; the budget curves are in Figure 2. It also shows the per-seed outcome of those co-learning experiments, the probes over time and window width, and episodes drawn on the grid. Throughout, an *arm* is one configuration of an experiment (for example, the follower with a prediction head), and a *experiment* is one experiment made of several arms. The figures add no new statistics: every number the paper claims is in the tables and their pages.

Figure 1 at full size. Its four panels are repeated below, one per figure, from the same data. Their reading is in that figure's caption.

In Figure 1(A), the history learner's gain lies on the identity line at values 0, 0.1, 0.3 and 0.5, the privileged learner's gain at the residual, and the prediction-head differences at zero; the mean at 0.5 falls slightly below the line because one seed collapsed, and the median lies on it. (B) Co-learning pairs lie above a partner-blind learner and the best blind patrol we found, and below the oracle that is shown the corner. (C) The objective raises decodability at step 3 by up to $+0.49$ (the larger step-1 gains of Table 5 use a different alignment); the only return difference near the margin is one toy setting, $+0.025$ $[-0.000, +0.074]$, driven by a single collapsed seed. (D) The registered $w = 8$ replication (Table 11); "predict" is the follower's next-action head, and "random" and "no loss" are the reading branch on random labels and without a prediction loss. The dashed line is a pair that never loses sight of its partner, from the legibility experiment at the same window; the legibility experiments at $w = 8$ and $w = 5$ are in Figure 10. Intervals are seed-level 95% bootstrap intervals; ten seeds per arm in the game, twenty in the toy and sixty in (D).

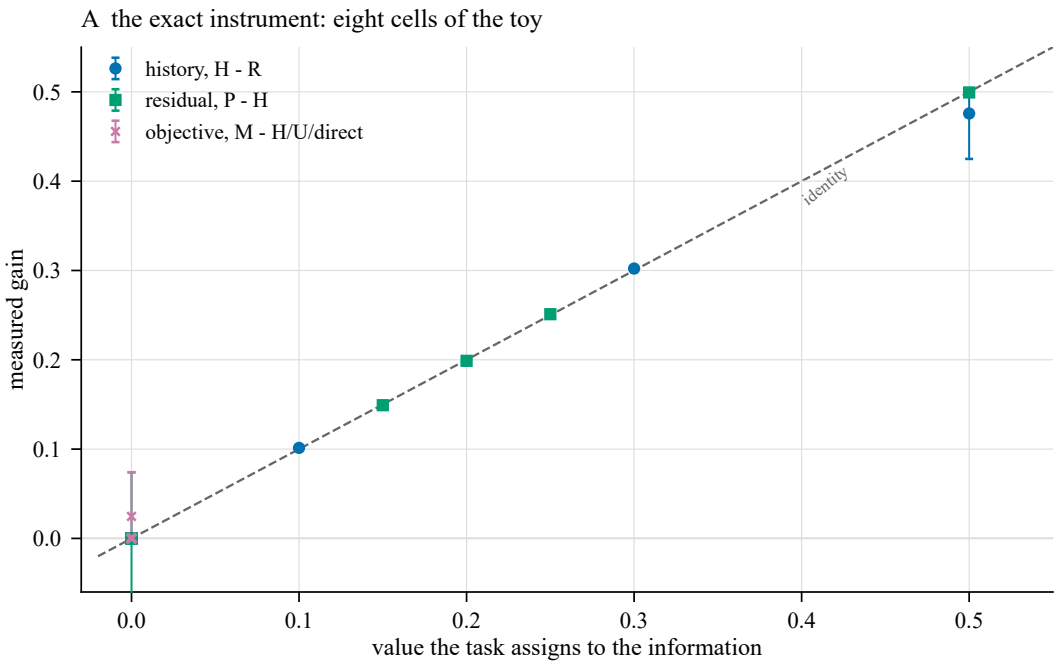

Figure 3: Panel A of Figure 1 at full size: the commitment toy, solved exactly at eight (g, p) settings (the toy's cells).

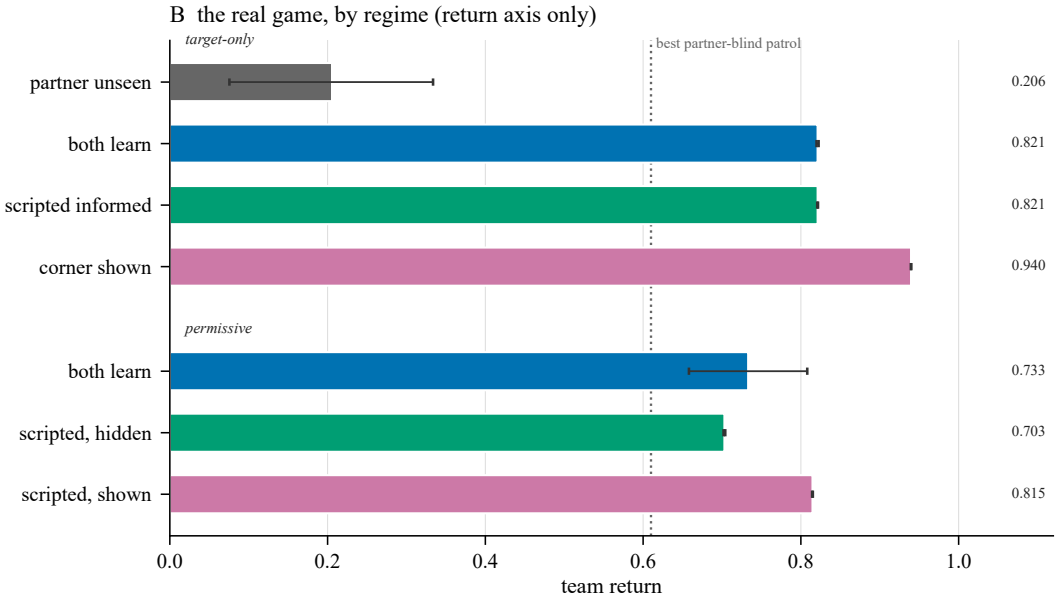

Figure 4: Panel B of Figure 1 at full size: Cooperative Reaching in each regime, on its own return axis.

In Figure 7(A), the four corners of the grid pay when both agents stand on one. Under the permissive rule two corners pay 1.0 and the other two pay 0.75. Under the target-only rule only the leader's private corner z pays, at 1.0; the other corners neither pay nor end the episode. The scripted leader is heuristic H07. It draws z uniformly and keeps it, and it stands still for a private delay of 0 to 6 steps. It then moves toward z, trying left, up, right and down in that order of priority. Each of its actions is replaced by a uniformly random one with probability 0.4. The sketch shows a leader walking from $(2, 1)$ to $(4, 4)$, with the noise omitted. The follower sees both positions and the leader's last executed

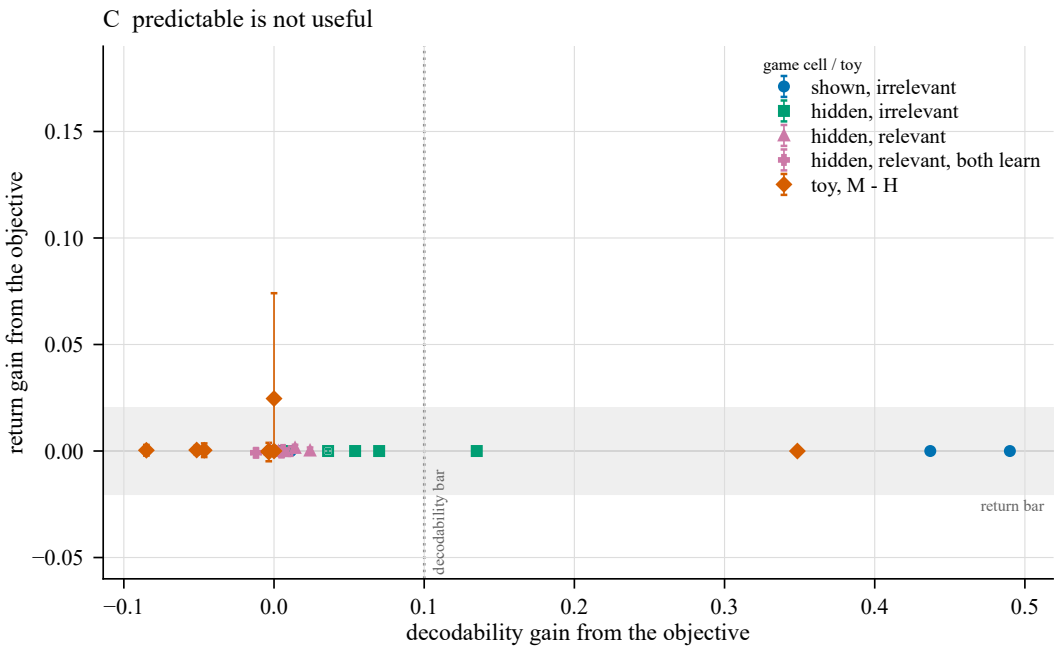

Figure 5: Panel C of Figure 1 at full size: the objective's gain in decodability plotted against its gain in return, with the two pre-registered margins.

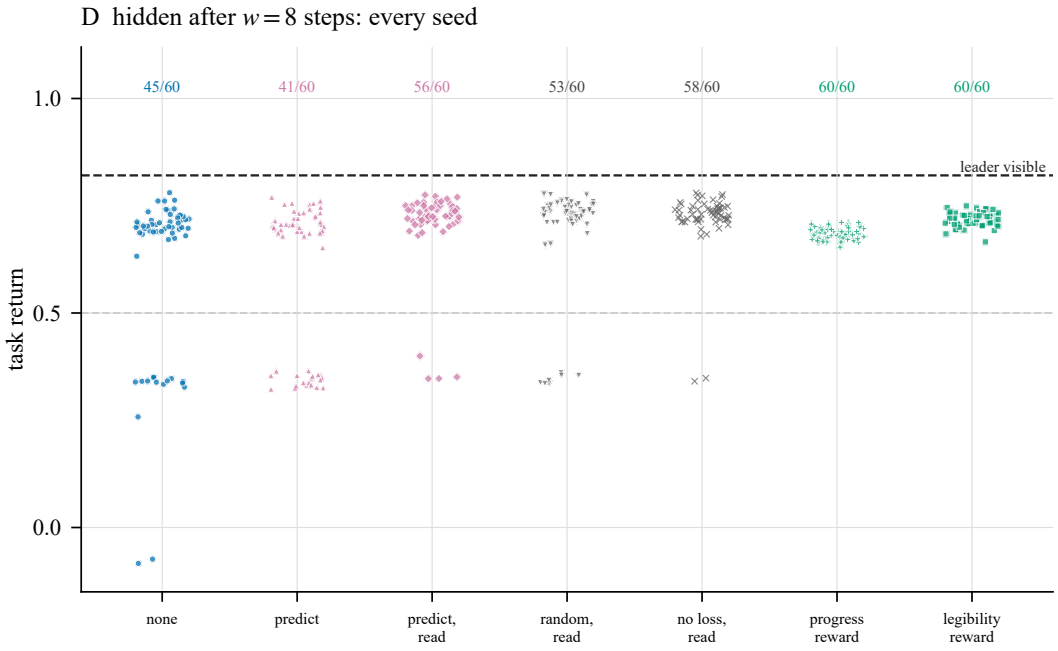

Figure 6: Panel D of Figure 1 at full size: every seed of the registered $w = 8$ replication, and the number of each arm's runs above the 0.5 return threshold.

action, but only for the leader's first w moves after the delay (drawn for $w = 3$; the experiments use $w = 3$, 5 and 8). After that it never sees the leader's cell, and in the hidden condition it never sees z. The follower pays 0.02 per step while it is off its rewarded corner, so waiting has a cost. At evaluation, the swap replaces the leader's inputs with those of another episode, so the leader's moves carry no information about this episode's z. The resulting drop in return measures how much the follower used those moves. The shown-corner and partner-blind arms control for the unfamiliarity

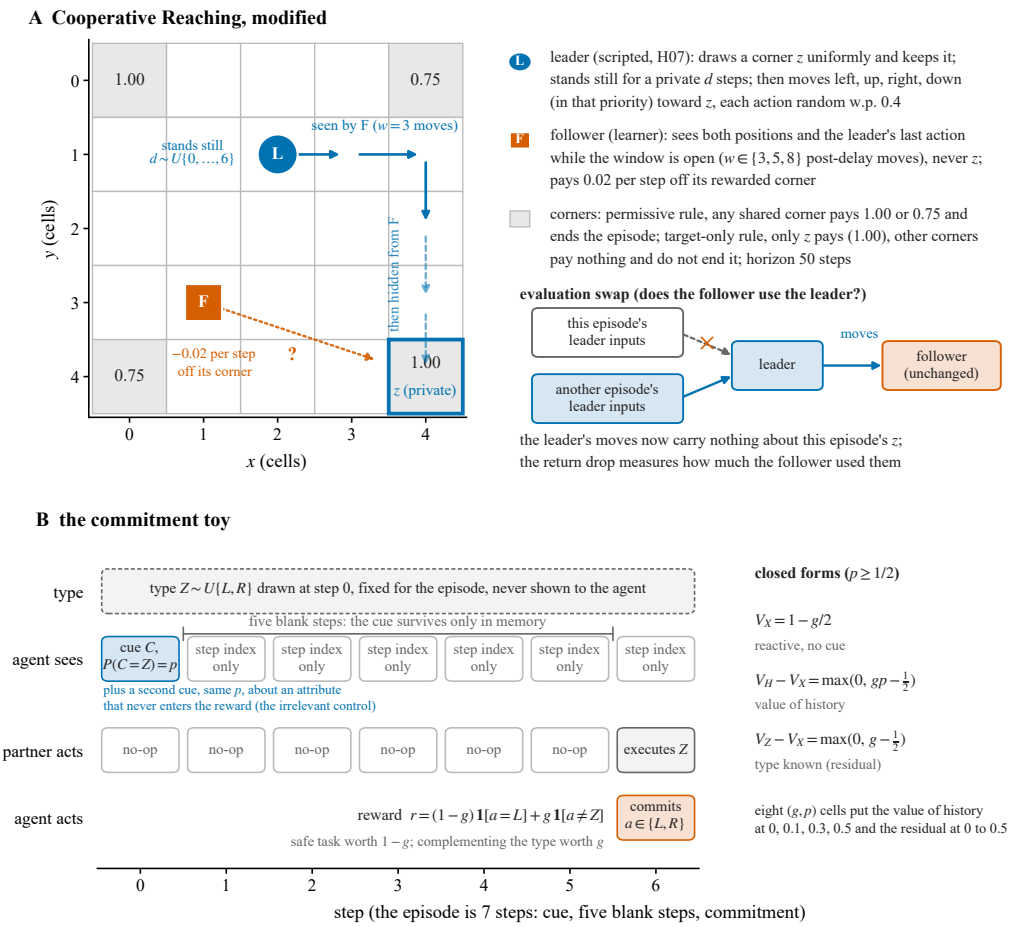

Figure 7: The two testbeds as schematics. (A) Modified Cooperative Reaching: a scripted or learning leader with a private corner and a follower that sees it only for a window of w moves. (B) The commitment task as a timeline: a cue, blank steps, one decision.

of swapped inputs. (B) A type $Z \in \{L, R\}$ is drawn uniformly and never shown. At step 0 the agent sees a cue C with $P(C = Z) = p$. It also sees a second cue of the same reliability about an attribute that never enters the reward. Steps 1 to 5 are blank, so the cue survives only in memory. At step 6 the agent commits to $a \in \{L, R\}$ and receives $(1 - g)\,\mathbf{1}[a = L] + g\,\mathbf{1}[a \neq Z]$. The safe action is worth $1 - g$, and complementing the type is worth g. The partner does nothing until that step, where it executes Z. The closed forms at the right give the value of history, $\max(0,\ gp - 1/2)$, and the residual, $\max(0,\ g - 1/2)$. The eight (g, p) cells place the value of history at 0, 0.1, 0.3 and 0.5 and the residual at 0 to 0.5. The experiments on these testbeds use 20 pre-registered seeds per arm in the toy (100 per arm in the collapse follow-up). The game uses 10 seeds per arm in the fixed-partner, positive-regime and scripted-retention experiments, and 60 per arm in the co-learning retention experiments. No seed is dropped.

In Figure 8(A), the fixed-partner 2×2 experiment crosses whether the corner is shown or hidden with whether it is relevant or irrelevant to the follower's reward, over 400 iterations. We call each of the four combinations a cell. Every arm converges to its cell's level: 0.815 in the three cells the corner does not gate and 0.703 in the hidden, relevant cell, within 0.003 of the no-objective arm. (B) The positive regime uses the target-only rule, with both agents learning. The head placements, the random-label control, the unshared pairs and the scripted informed leader all reach the no-objective pair's 0.821 along the same path. The corner-shown reference reaches 0.940, and the partner-blind follower stays near 0.2. (C) In the permissive-rule experiment, every co-learning pair converges to the convention level whatever its objective. This holds with the corner hidden or shown, and with shared or unshared parameters. That level is 0.71 to 0.76, the mean over seeds of a bimodal

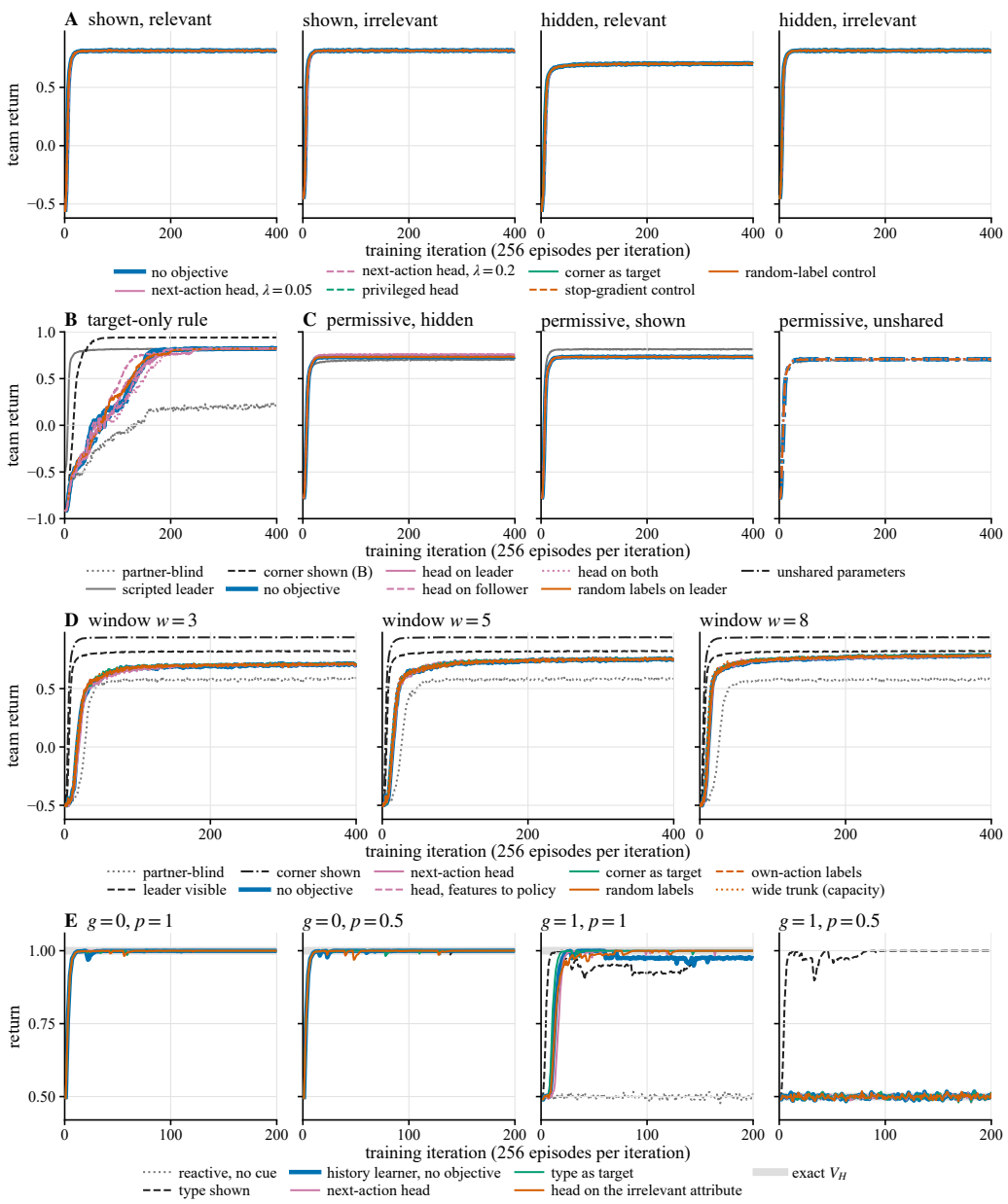

Figure 8: Learning curves of the experiments with a scripted partner and of the positive regime (10 seeds per arm in the game, A to D; 20 in the toy, E). The no-objective arm is the thick blue line. In these experiments every objective and control follows the no-objective curve throughout training.

outcome: 0.83 for pairs that settle on a corner paying 1.0, and 0.58 for pairs on a 0.75 corner. The scripted references sit at 0.703 and 0.815. (D) The retention experiment uses the scripted leader at windows 3, 5 and 8. Every objective and control arm lies on the no-objective curve (0.708, 0.752, 0.781). These curves sit between the partner-blind floor (0.58) and the references with the leader visible (0.822) and the corner shown (0.940). (E) The commitment toy runs for 200 iterations. The history learner reaches the exact value V_H (pale band) in every cell, and the objective arms lie on it. The reactive learner stays at 0.5 where the type matters. Where the cue is uninformative, only the privileged arm reaches 1.0. The no-objective and privileged means at $g = 1$, $p = 1$ sit at 0.975, because one seed of each collapsed to 0.5 after reaching the ceiling (Appendix A). Takeaway: in every experiment shown, the prediction objective and its controls trace the no-objective curve from

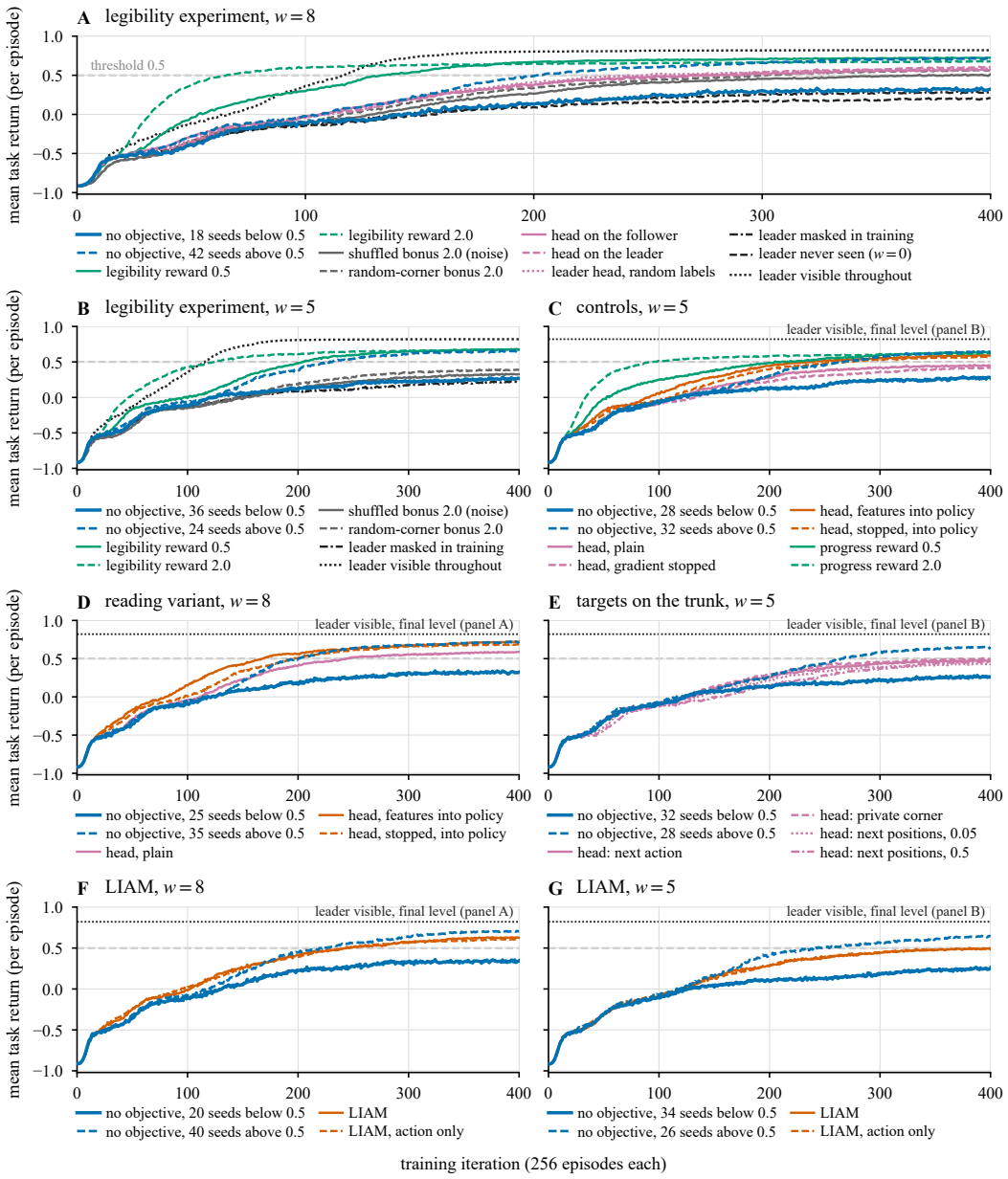

Figure 9: Learning curves of the seven 400-iteration co-learning experiments (60 seeds per arm). The leader rewards, and at $w = 8$ the head whose features the policy reads, raise mean return above the no-objective arm; most other arms stay close to it.

the first iteration to the last. This is consistent with the flat return contrasts of the tables not being an artefact of the training budget used.

In Figure 9, the plotted return is the sparse task return throughout, so the legibility and progress arms are drawn without their bonus. Colours: blue is no objective. Green is a reward paid to the team for the leader's behaviour (legibility, or the matched progress control). Grey is a content-free reward control (the bonus of another episode, or a bonus for a corner the reward never uses). Purple is a prediction head whose features the policy does not read. Vermilion is modelling features that enter the policy: the head read by the policy, with and without its gradient into the shared trunk, and the LIAM-class separate encoder. Black is what the follower sees: the leader visible throughout, masked in training, or never seen. The no-objective arm of each experiment is split into two groups

by final return, the mean over the last tenth of training. One group holds the seeds that end above the 0.5 threshold (grey dashed line); the other holds those that end below it. The legend gives the count in each group. The outcome is bimodal. Averaged within each experiment, the below-threshold seeds end at 0.24 to 0.34 and the above-threshold seeds at 0.64 to 0.71; individual below-threshold seeds can end lower (Figure 10). Panels C to G have no leader-visible arm. In those panels the dotted line marks that reference's final level in the legibility experiment at the same window (panel A or B). In panel E the next-positions head predicts the leader's next three grid positions at the stated weight. Takeaway: at both weights the reward arms pull most seeds above the threshold, and so does the head whose features the policy reads at $w = 8$. Their curves leave the no-objective mean within the first 150 iterations and end near the level of the above-threshold no-objective seeds. At $w = 5$ the same feature-reading arms (panel C) lift the mean by about 0.1 and sit between the two groups. The content-free controls, the heads whose features the policy does not read, and the LIAM-class encoder under its stop-gradient stay within 0.1 of the no-objective mean. The leader-visible reference sits above every arm in every experiment.

The panels of Figure 10 are (A) the legibility experiment at $w = 8$; (B) the legibility experiment at $w = 5$; (C) the controls experiment at $w = 5$; (D) the feature-reading experiment at $w = 8$; (E) the prediction-target experiment at $w = 5$; (F, G) the LIAM-class experiments at $w = 8$ and $w = 5$. Labels: "sg" is a head whose gradient is stopped at the shared trunk. "read" is a head whose features enter the policy. "motion" is a head predicting the leader's next three positions, at weight 0.05 or 0.5. "private corner" uses the leader's corner as a privileged label. "LIAM, action only" is the encoder without its observation head. The two blind arms train the follower without its view of the leader, with and without the window flag. The outcome is bimodal in every experiment. Seeds settle either in an upper group, at about 0.55 to 0.75, or in a lower group, at about 0.2 to 0.35, with almost nothing in between. The upper group is consistent with a follower that reads and keeps the corner (Figure 12). The no-objective arm splits between the groups in every experiment (24 to 42 of 60 above the threshold). Two kinds of arm shrink the lower group the most. The first is the two shaping rewards on the leader, legibility and progress, which place 57 to 60 of 60 seeds above the threshold at both weights. The second is the head whose features the policy reads: 58 and 53 of 60 at $w = 8$, and 46 and 42 of 60 at $w = 5$. The other arms leave the lower group in place (22 to 47 of 60). These are the plain head, its stop-gradient and target variants, the content-free controls, and the LIAM-class encoder under its stop-gradient. The same encoder without the stop-gradient reaches 47 of 60, with a benefit verdict (Table 9). The $w = 5$ LIAM half ran on a second hardware platform; every other experiment ran on the first platform.

In Figure 11(A), windows are $w = 3, 5, 8$. Lines are the mean over 10 seeds per arm, with shaded seed-level 95% bootstrap bands. The star is the computed accuracy of one declaration at the end of the window, 0.63, 0.79 and 0.91, which serves as the ceiling; chance is 0.25. When the window ends the follower holds the corner at 0.58, 0.74 and 0.79, within 0.05 to 0.12 of the computed ceiling (Section 6.3 compares with the best accuracy on the same episodes instead, which leaves gaps of 0.03 to 0.07). It loses part of the corner while it waits, falling to between 0.39 and 0.49 ten steps later. The head's curves lie on the no-objective curves; in the tables, its effect on later decodability is inconclusive at ten seeds (Table 6). (B) The legibility probe $L(k)$ is the held-out accuracy of a probe reading the leader's corner from its first k moves. It is shown for the co-learning legibility experiments at $w = 8$ (left) and $w = 5$ (right), as the mean over 60 seeds per arm with bootstrap bands. The dotted line is the no-objective curve plus the pre-registered margin of 0.10. The dashed vertical line marks the $w - 1$ moves that a follower at window w can see. At both weights the legibility reward raises the leader's legibility at every k. At $w = 8$ the gain is 0.045 to 0.081, inside the margin. At $w = 5$ it is 0.08 to 0.13, and it reaches the margin only from $k = 4$ on. The reward-noise and wrong-corner controls fall below the no-objective arm, and the follower-side prediction head tracks the no-objective curve. (C) The follower's commit step is measured relative to the end of the window. Each value is one seed (the mean over that seed's test episodes). Boxes span the quartiles and whiskers the 5th to 95th percentile. The median is a thick bar; the mean is a thin bar and is printed at the right. There are 60 seeds per arm, and each experiment is compared with its own no-objective arm. At $w = 5$ the legibility reward brings the commit forward at both weights (means 6.2 and 6.7 against 10.0). At $w = 8$ it does so at one weight only (means 2.1 and 4.2 against 3.4). The head whose features the policy reads also commits earlier (1.3 and 1.8 against 3.9). The plain next-action head and the follower-side prediction head do not bring the commit forward by a comparable amount (3.4 against 3.9; 3.7 against 3.4), and both controls commit later.

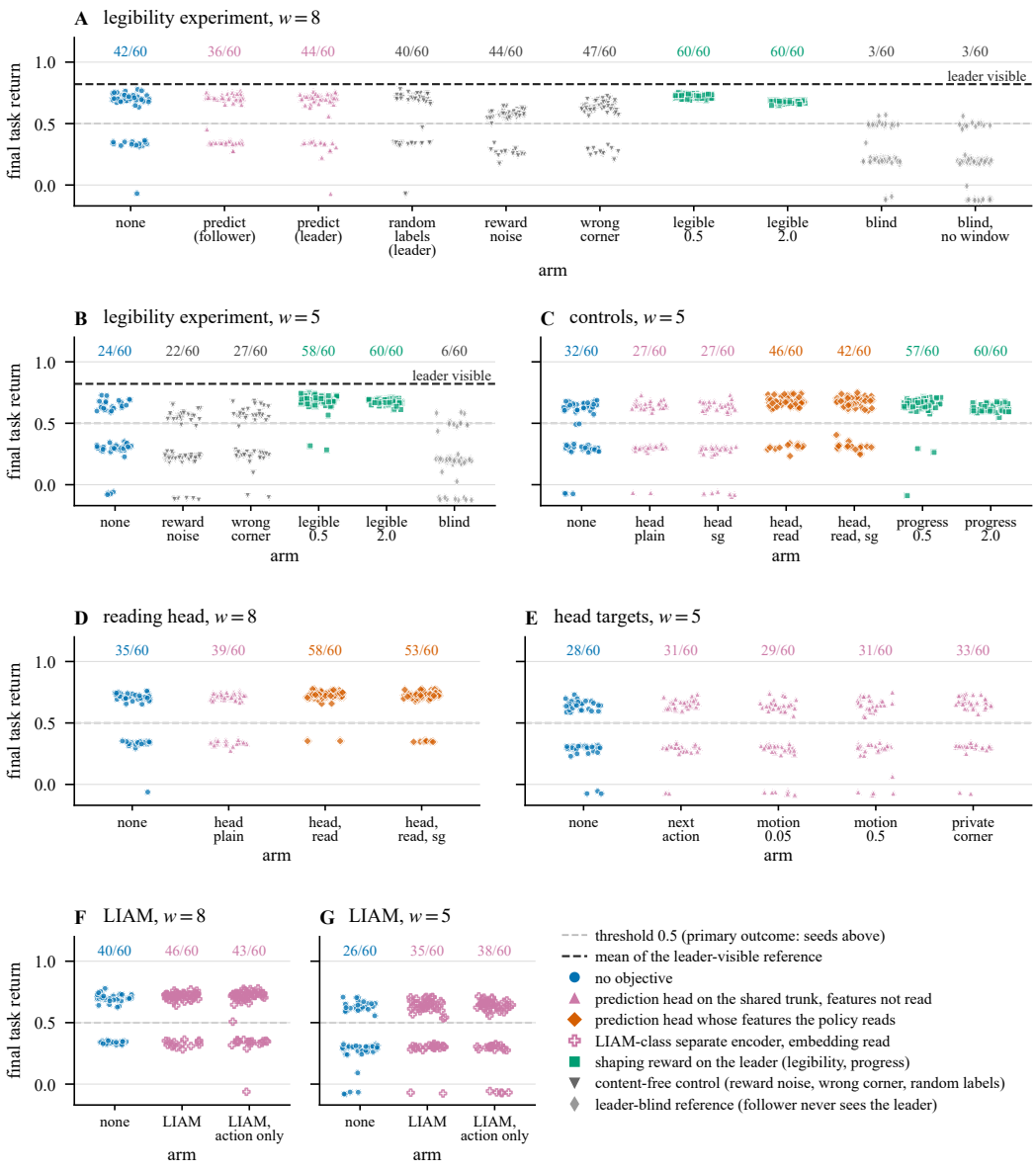

Figure 10: Every seed of the seven 400-iteration co-learning experiments: final task return per seed, with the count above the 0.5 threshold printed over each strip. Outcomes are bimodal in every experiment; the leader rewards and the head read by the policy move the most seeds into the upper group.

In each row of Figure 12, the four grid panels show one evaluation episode each, chosen without looking at them. They are the first four episodes of the archive whose leader delay d is at most 3, so that the window closes by step 8. Cell $(0, 0)$ is at the top left; x is the column and y the row. The leader's private corner is outlined in blue. The leader's path (blue) is drawn from the end of its delay and the follower's path (vermilion) from step 0; both darken with time. Hollow marks are the start cells. The filled circle is the leader's cell at the last step at which the follower sees it ($t = d + 4$). The filled square is the follower's cell when the window ends ($t = d + 5$). Light circles mark cells in which an agent stood still for two or more steps, with area proportional to the number of steps. A star marks the pair's arrival on the leader's corner, and crosses mark each agent's final cell when the pair never arrives. The wider panels E to H use all 256 episodes of the same checkpoint. The black

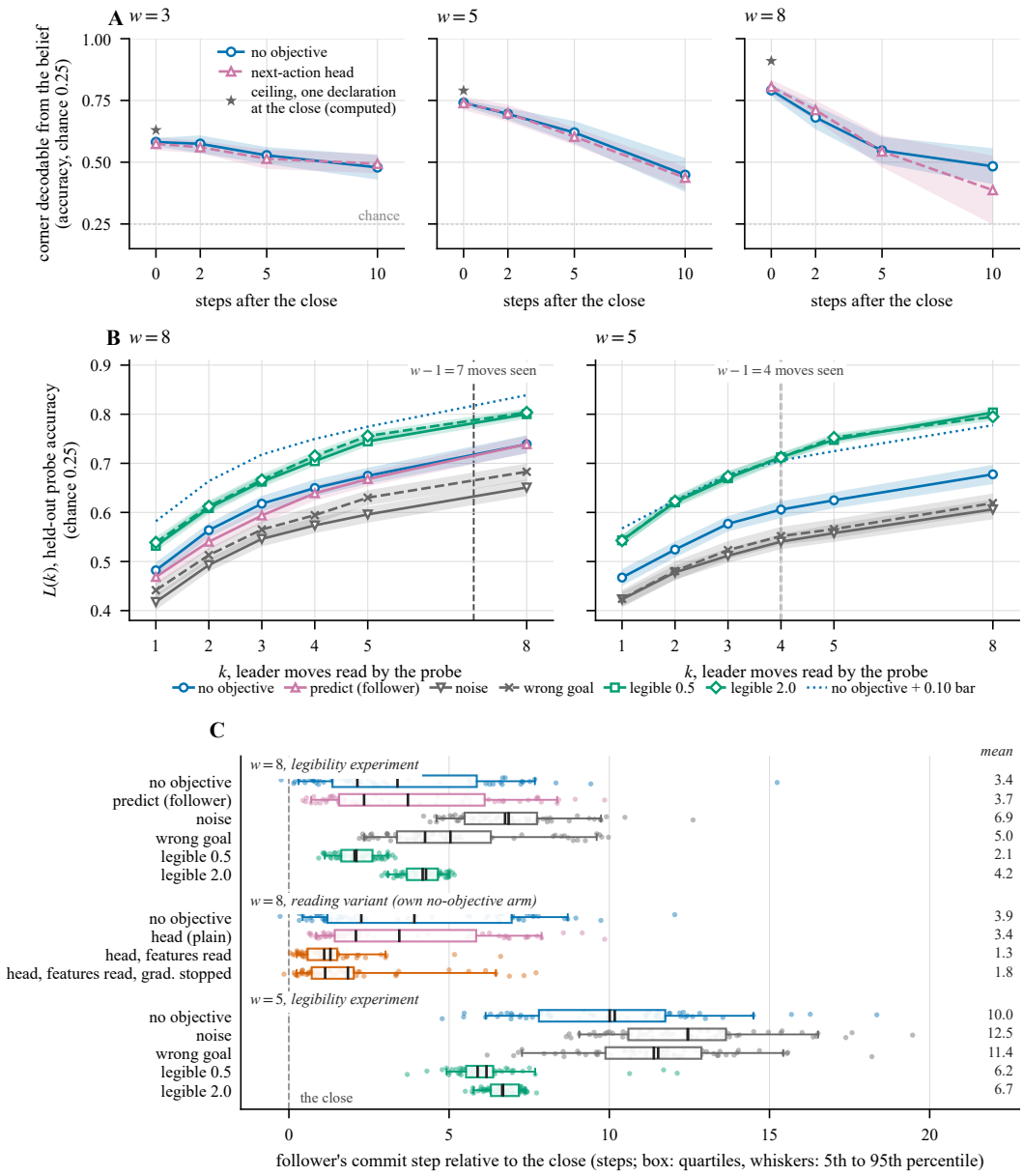

Figure 11: What the probes see. (A) Decodability of the corner from the follower's memory after the window closes, with the scripted leader. (B) Legibility of the co-learning leader from its first k moves. (C) The co-learning follower's commit step relative to the end of the window. The head's curve in (A) lies on the no-objective curve.

curve is the fraction of episodes in which the pair stands on the leader's corner by step t; the episode is then paid and ends. The dashed curve is the fraction in which the follower alone has stood on it by step t. The other rows' black curves are repeated in grey, and the grey band covers the steps at which the window can end. The printed mean return is over these 256 episodes; the final returns in the caption are the Final-k values of the run summaries. Each leader action is replaced by a random one with probability 0.4, so even a goal-directed leader's path is jagged. Early in training (A) the follower goes to $(0, 0)$ in every episode and waits there. The pair is then paid only when the leader's corner happens to be $(0, 0)$ (0.25 of the episodes). In the below-threshold run (B) the follower has a fixed left-column route: $(0, 0)$ first in 218 of 256 episodes and $(0, 4)$ first in the rest. It never visits a right-hand corner. The pair arrives in every episode whose corner is $(0, 0)$ or $(0, 4)$ and in

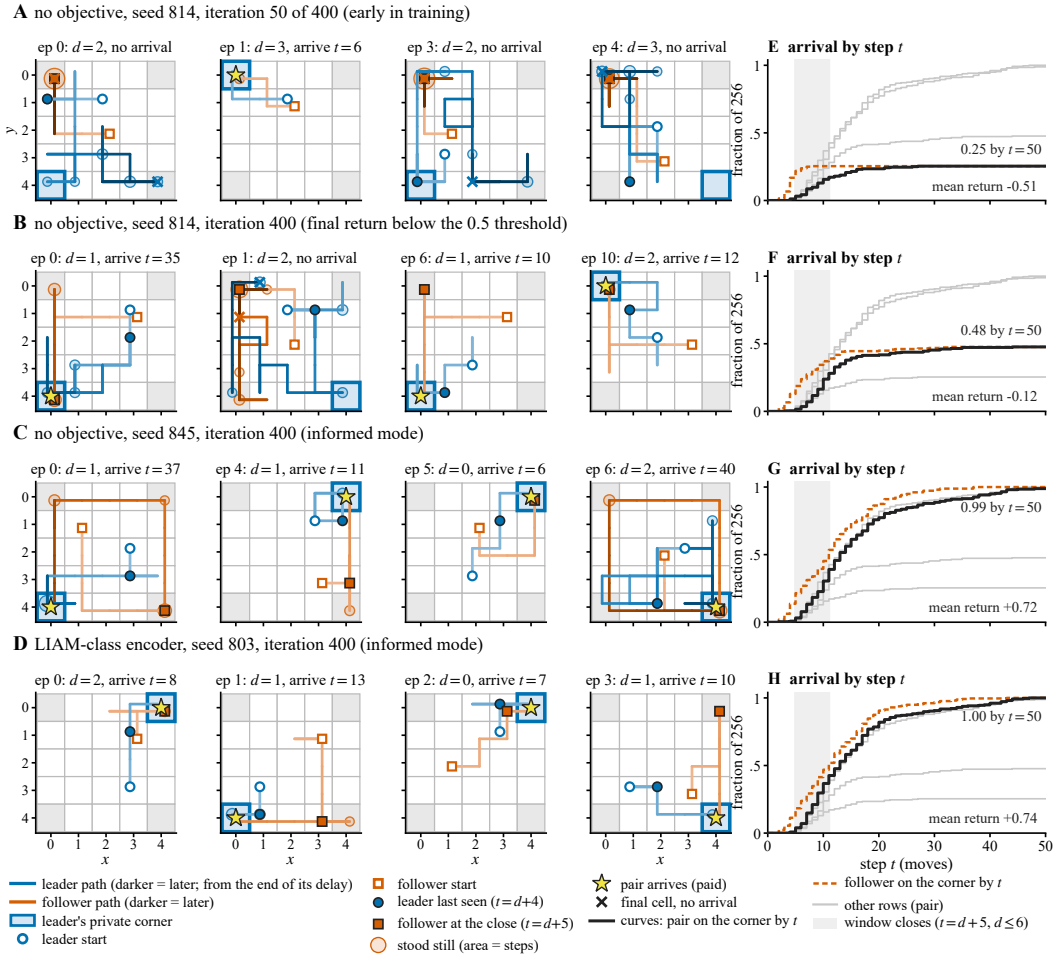

Figure 12: Four runs of the $w = 5$ co-learning experiment drawn on the grid: a no-objective run early and late in training (A, B; below the threshold), and two runs above it (C, no objective; D, LIAM-style encoder). The below-threshold run follows a fixed route over part of the corners; the others reach the leader's corner.

none of the others (0.48 in all). This looks like a convention rather than use of the leader. In the informed runs (C, D) the pair arrives in 0.99 and 1.00 of the episodes, most of them after the window ends (0.83 and 0.81). The follower often visits one convention corner first ($(4, 4)$ in C, $(4, 0)$ in D) and then heads to the leader's corner. Its mean distance to that corner falls from 2.8 cells when the window ends to below 0.1 at the end of the episode. In these plots we could not tell the no-objective and LIAM-class informed runs apart. We also checked every run of the experiment, using the traces of the last checkpoint. All 81 runs that end below the threshold show the signature of run B. They arrive in over 90% of the episodes whose corner lies on the follower's route and in under 10% of the others (about 70% of episodes in all, two or three corners). None of the 99 informed runs shows it; they arrive in 95 to 96% of episodes, at all four corners. In this experiment, then, runs that stay below the threshold are consistent with a fixed convention over part of the corners rather than a follower that wanders.

Table 9: The remaining prediction arms, each experiment against its own no-objective arm on the same seeds (60 seeds per arm, 400 iterations, parameter-shared self-play). Top: runs with task return above 0.5, with the rate difference and its Newcombe interval, and the return difference with its seed-level 95% interval and verdict at the 0.02 margin (B benefit, I inconclusive, E equivalent, D deficit), with Holm correction within each experiment. Bottom: the follower's decodability of the corner at the end of the window and its commit step relative to it (margin 0.10). The head whose features the policy reads clears both margins at $w = 8$ and improves the follower's decodability and commitment. Three prediction targets on the shared network (including the motion target of BEPAL at two weights and the private corner as a privileged label) and a LIAM-style separate encoder at both windows are inconclusive. Without its gradient block, so that the task loss shapes the encoder, the LIAM-style arm clears the return margin ($+0.098$) and misses the 0.2 rate margin ($+0.18$). The $w = 5$ LIAM half ran on a second hardware platform, every other experiment on the first platform (Appendix H.1).

arm	above 0.5 (rate minus ref)	return (verdict)
The reading variant at $w = 8$ (own no-objective arm)		
no objective (level)	35 of 60	0.549
head on the follower (plain)	39 ($+0.07$ $[-0.11, +0.23]$)	$+0.034$ $[-0.034, +0.103]$ I
head, features into the policy	58 ($+0.38$ $[+0.24, +0.51]$)	$+0.167$ $[+0.113, +0.221]$ B
head, gradient stopped, features into the policy	53 ($+0.30$ $[+0.14, +0.44]$)	$+0.134$ $[+0.076, +0.194]$ B
LIAM-class encoder at $w = 8$ (own no-objective arm)		
no objective (level)	40 of 60	0.581
LIAM (observation and action)	46 ($+0.10$ $[-0.06, +0.25]$)	$+0.045$ $[-0.018, +0.105]$ I
LIAM, action only	43 ($+0.05$ $[-0.11, +0.21]$)	$+0.023$ $[-0.043, +0.088]$ I
LIAM-class encoder at $w = 5$ (against its own no-objective arm)		
no objective (level)	26 of 60	0.412
LIAM (observation and action)	35 ($+0.15$ $[-0.03, +0.32]$)	$+0.077$ $[+0.001, +0.149]$ I
LIAM, action only	38 ($+0.20$ $[+0.02, +0.36]$)	$+0.076$ $[-0.006, +0.154]$ I
Targets on the shared trunk at $w = 5$ (own no-objective arm)		
no objective (level)	28 of 60	0.439
next action ($\lambda = 0.05$)	31 ($+0.05$ $[-0.13, +0.22]$)	$+0.030$ $[-0.044, +0.103]$ I
next three positions (0.05)	29 ($+0.02$ $[-0.16, +0.19]$)	-0.011 $[-0.091, +0.069]$ I
next three positions (0.5)	31 ($+0.05$ $[-0.13, +0.22]$)	$+0.015$ $[-0.065, +0.096]$ I
the private corner (0.05, privileged)	33 ($+0.08$ $[-0.09, +0.25]$)	$+0.054$ $[-0.020, +0.130]$ I
LIAM-class encoder without its stop-gradient at $w = 8$ (own no-objective arm)		
no objective (level)	36 of 60	0.552
LIAM, embedding shaped by the task loss	47 ($+0.18$ $[+0.02, +0.34]$)	$+0.098$ $[+0.032, +0.163]$ B

arm	decodable at window end	commit relative to window end
The reading variant at $w = 8$ (own no-objective arm)		
head on the follower (plain)	$+0.008$ $[-0.028, +0.044]$	-0.47 $[-1.53, +0.55]$
head, features into the policy	$+0.078$ $[+0.047, +0.109]$	-2.62 $[-3.48, -1.80]$
head, gradient stopped, features into the policy	$+0.051$ $[+0.016, +0.086]$	-2.07 $[-3.00, -1.17]$
LIAM-class encoder at $w = 8$ (own no-objective arm)		
LIAM (observation and action)	$+0.021$ $[-0.015, +0.057]$	-0.74 $[-1.69, +0.25]$
LIAM, action only	$+0.023$ $[-0.012, +0.058]$	-0.33 $[-1.33, +0.72]$
LIAM-class encoder at $w = 5$ (against its own no-objective arm)		
LIAM (observation and action)	-0.005 $[-0.032, +0.023]$	-0.31 $[-1.22, +0.60]$
LIAM, action only	-0.007 $[-0.038, +0.023]$	-0.26 $[-1.28, +0.78]$
Targets on the shared trunk at $w = 5$ (own no-objective arm)		
next action ($\lambda = 0.05$)	$+0.007$ $[-0.023, +0.036]$	-0.55 $[-1.46, +0.36]$
next three positions (0.05)	$+0.019$ $[-0.010, +0.047]$	-0.06 $[-1.09, +0.98]$
next three positions (0.5)	$+0.043$ $[+0.011, +0.074]$	-0.75 $[-1.78, +0.30]$
the private corner (0.05, privileged)	$+0.058$ $[+0.029, +0.088]$	-1.03 $[-2.01, -0.06]$
LIAM-class encoder without its stop-gradient at $w = 8$ (own no-objective arm)		
LIAM, embedding shaped by the task loss	$+0.028$ $[-0.006, +0.062]$	-1.42 $[-2.36, -0.47]$

Table 10: The three robustness experiments registered in advance (60 seeds per arm, seeds 800 to 859, parameter-shared self-play unless stated). Top: the same pair trained for 1600 iterations, four times the budget, at each window, beside its 400-iteration experiment on the same seeds. The change in the count above 0.5 carries a Newcombe interval, and the reward's return difference is given at each budget with its seed-level 95% interval and verdict at the 0.02 margin (B benefit, I inconclusive, D deficit). Without an objective, 57 of 60 pairs at $w = 5$ and 60 of 60 at $w = 8$ pass the threshold by 1600 iterations, and the reward's advantage falls from $+0.260$ to $+0.003$ and from $+0.131$ to -0.038: the runs below the threshold are slow rather than stuck, and the reward makes pairs reach the informed behaviour sooner, not better. Bottom: one network per agent. The two rewards and the head whose features the policy reads keep their effects against the unshared no-objective pair, which itself passes the threshold more often than the shared pairs (40 against 24 to 32 of 60). At the end of the window the follower decodes the corner better under both rewards ($+0.150$ and $+0.137$, clearing the 0.10 margin) and within the margin under the reading head ($+0.069$), and commits 1.8 to 2.9 steps earlier under all three, as in the shared experiments.

Four times the budget (same seeds at 400 and 1600 iterations)

	400 iterations	1600 iterations
$w = 5$ *(400 iterations from the replication experiment)*		
no objective, runs above 0.5	24 of 60	57 of 60 ($+0.55$ [$+0.40$, $+0.67$])
legibility reward $\beta = 0.5$, runs above 0.5	58 of 60	60 of 60 ($+0.03$ [-0.03, $+0.11$])
no objective, mean return	0.415	0.739
legibility reward $\beta = 0.5$, mean return	0.675	0.742
reward minus no objective (verdict)	$+0.260$ [$+0.203$, $+0.317$] B	$+0.003$ [-0.017, $+0.028$] I
$w = 8$ *(400 iterations from the main experiment)*		
no objective, runs above 0.5	42 of 60	60 of 60 ($+0.30$ [$+0.18$, $+0.43$])
legibility reward $\beta = 0.5$, runs above 0.5	60 of 60	60 of 60 ($+0.00$ [-0.06, $+0.06$])
no objective, mean return	0.591	0.788
legibility reward $\beta = 0.5$, mean return	0.721	0.751
reward minus no objective (verdict)	$+0.131$ [$+0.085$, $+0.181$] B	-0.038 [-0.043, -0.033] D

Unshared parameters, $w = 5$ (one actor network per agent; own no-objective arm)

arm	above 0.5 (rate minus ref)	return (verdict)
no objective (level)	40 of 60	0.530
legibility reward $\beta = 0.5$	60 ($+0.33$ [$+0.21$, $+0.46$])	$+0.160$ [$+0.118$, $+0.203$] B
progress reward $\beta = 0.5$	60 ($+0.33$ [$+0.21$, $+0.46$])	$+0.165$ [$+0.123$, $+0.208$] B
head, features into the policy	54 ($+0.23$ [$+0.09$, $+0.37$])	$+0.110$ [$+0.057$, $+0.161$] B

arm (unshared)	decodable at window end	commit relative to window end
legibility reward $\beta = 0.5$	$+0.150$ [$+0.128$, $+0.173$]	-2.91 [-3.51, -2.34]
progress reward $\beta = 0.5$	$+0.137$ [$+0.114$, $+0.161$]	-2.43 [-3.02, -1.87]
head, features into the policy	$+0.069$ [$+0.041$, $+0.098$]	-1.83 [-2.55, -1.11]

Table 11: A fresh replication at $w = 8$ registered in advance (seeds 900 to 959, never used before; 60 per arm, 420 runs, one code version, parameter-shared self-play, 400 iterations). Top: runs with task return above 0.5, with the rate difference from no objective and its Newcombe interval, and the return difference with its seed-level 95% interval and verdict at the 0.02 margin (B benefit, I inconclusive), Holm across the six arms. Bottom: the feature-reading head against the two controls fixed in advance. The ordering of the earlier experiments replicates: the plain head is inconclusive, and the reading head and both rewards help. But the same reading branch helps as much when trained on random labels or with no prediction loss at all, so, as the registration required, the gain is attributed to an extra input branch that the policy reads and the task loss shapes, not to partner content.

arm	above 0.5 (rate minus ref)	return (verdict)
no objective (level)	45 of 60	0.602
next-action head, features not read	41 (-0.07 $[-0.22, +0.09]$)	-0.006 $[-0.072, +0.062]$ I
next-action head, features read by the policy	56 ($+0.18$ $[+0.05, +0.31]$)	$+0.105$ $[+0.052, +0.164]$ B
the same reading branch, random labels	53 ($+0.13$ $[-0.01, +0.27]$)	$+0.086$ $[+0.028, +0.148]$ B
the same reading branch, no prediction loss	58 ($+0.22$ $[+0.09, +0.34]$)	$+0.118$ $[+0.067, +0.175]$ B
progress reward, $\beta = 0.5$	60 ($+0.25$ $[+0.14, +0.37]$)	$+0.083$ $[+0.036, +0.137]$ B
legibility reward, $\beta = 0.5$	60 ($+0.25$ $[+0.14, +0.37]$)	$+0.115$ $[+0.068, +0.170]$ B
reading head minus the random-label branch		$+0.018$ $[-0.021, +0.061]$ I
reading head minus the loss-free branch		-0.013 $[-0.044, +0.017]$ I

Table 12: A capacity control at $w = 8$ registered in advance (fresh seeds 1100 to 1159, 60 per arm, 240 runs, one code version, parameter-shared self-play, 400 iterations). Parameters are the trainable parameters on the policy path added to the no-objective network. Top: runs with task return above 0.5, with the rate difference from no objective and its Newcombe interval, and the return difference with its seed-level 95% interval and verdict at the 0.02 margin (B benefit, I inconclusive), Holm across the four registered contrasts. Bottom: the reading branch against the parameter-matched trunk. A trunk widened to the branch's parameter count does not resolve a gain; the branch is a benefit over both, and ten times the added parameters recover only part of it.

arm	above 0.5 (rate minus ref)	return (verdict)
no objective (level)	37 of 60	0.560 (mean)
reading branch, no prediction loss (+8,576 parameters)	58 ($+0.35$ $[+0.21, +0.48]$)	$+0.163$ $[+0.111, +0.216]$ B
no objective, trunk widened to 136 units (+9,992)	43 ($+0.10$ $[-0.07, +0.26]$)	$+0.036$ $[-0.039, +0.108]$ I
no objective, trunk widened to 192 units (+90,688)	49 ($+0.20$ $[+0.04, +0.35]$)	$+0.094$ $[+0.031, +0.158]$ B
reading branch minus the matched trunk		$+0.127$ $[+0.074, +0.185]$ B

Table 13: Followers trained against a population of frozen leaders at $w = 8$ (registered in advance; 20 seeds per arm, 120 runs). Each pool holds the final leaders of replication seeds 900 to 929 of one arm (Table 11); seeds 930 to 959 are held out and meet each follower only at evaluation. Returns are means over 256 sampled episodes per run: with the training pool, with a held-out leader of the same arm, and with the leader's inputs swapped for another episode's. Bottom: held-out return differences with seed-level 95% intervals and verdicts at the 0.02 margin (B benefit, E equivalent, I inconclusive), Holm within each registered family.

leader pool	follower	training pool	held-out leader	above 0.5	swapped
no objective	no objective	0.637	0.522	12 of 20	0.092
no objective	reading branch	0.643	0.534	12 of 20	0.079
progress reward	no objective	0.717	0.707	20 of 20	0.216
progress reward	reading branch	0.728	0.718	20 of 20	0.223
legibility reward	no objective	0.715	0.714	20 of 20	-0.042
legibility reward	reading branch	0.729	0.716	20 of 20	-0.038
no-objective followers, progress reward pool minus no-objective pool				$+0.185$ $[+0.112, +0.262]$ B	
no-objective followers, legibility reward pool minus no-objective pool				$+0.193$ $[+0.121, +0.270]$ B	
reading followers, progress reward pool minus no-objective pool				$+0.184$ $[+0.111, +0.263]$ B	
reading followers, legibility reward pool minus no-objective pool				$+0.182$ $[+0.110, +0.260]$ B	
reading minus no-objective follower, no-objective pool				$+0.013$ $[-0.093, +0.120]$ I	
reading minus no-objective follower, progress reward pool				$+0.012$ $[-0.005, +0.028]$ I	
reading minus no-objective follower, legibility reward pool				$+0.002$ $[-0.009, +0.012]$ E	

Table 14: Cross-play in the $w = 8$ replication: each trained follower is paired with the frozen leader of a different seed (seed k with seed $k + 1$ of the leader's arm; 60 pairings per cell, 256 sampled episodes each). Cells give the mean return and, in parentheses, the pairings above 0.5. Bottom: differences with seed-level 95% intervals and verdicts at the 0.02 margin (B benefit, D deficit, I inconclusive), Holm within each registered family. These comparisons are descriptive.

follower \ leader	no objective	progress reward	legibility reward	reading head
no objective	0.369 (22)	0.517 (42)	0.443 (28)	0.430 (28)
next-action head	0.356 (22)	0.488 (34)	0.407 (24)	0.404 (24)
reading head	0.460 (33)	0.596 (54)	0.525 (42)	0.507 (40)
progress reward	0.477 (36)	0.620 (58)	0.553 (44)	0.537 (44)
legibility reward	0.447 (26)	0.560 (50)	0.658 (57)	0.484 (34)

no-objective follower, progress leader minus no-objective leader	$+0.149$ $[+0.073, +0.227]$ B
no-objective follower, legibility leader minus no-objective leader	$+0.075$ $[-0.001, +0.154]$ I
progress follower, no-objective leader minus progress leader	-0.142 $[-0.185, -0.104]$ D
legibility follower, no-objective leader minus legibility leader	-0.211 $[-0.257, -0.168]$ D

Table 15: OvercookedV2 with a frozen full-view partner and a learner that sees two cells around itself (20 seeds per arm, memory trained through time). Final sparse return over the last tenth of training; differences from no objective with seed-level 95% intervals and the verdict at the margin of 5 (E equivalent, I inconclusive), Holm across the four arms; and the number of runs in which the learner loads pots.

arm	final return	minus no objective (verdict)	loading role
no objective	204.5		18 of 20
next-action head	201.6	-2.91 $[-6.38, +0.53]$ I	13 of 20
next-action head, random labels	205.4	$+0.91$ $[-1.54, +3.50]$ E	19 of 20
next-action head, features read by the policy	201.9	-2.60 $[-6.62, +1.41]$ I	13 of 20
recipe as the target (privileged label)	205.2	$+0.74$ $[-2.22, +3.60]$ E	18 of 20