TEST-TIME POLICY IMPROVEMENT FOR COORDINATION WITH UNSEEN PARTNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cooperative multi-agent reinforcement learning prepares an agent for unseen partners by training it against a diverse population, so that one policy, fixed before deployment, generalises to partners it never trained with. But a policy committed before the partner is met must compromise whenever plausible partners call for incompatible responses, a gap no training-time diversity can close. We address this gap by introducing test-time best response (TTBR): after a short observation phase (16 episodes per seat), the agent forms an exact posterior over an explicit class of partner programs and policies and improves its own policy against it in the simulator. We show that the regret of the exact posterior response is bounded by the conflict among candidate partners, the cross-best-response regret that underlies minimum coverage sets, times the residual posterior odds, a guarantee that also predicts when partner identification cannot help. The resulting algorithm (i) responds to the partner actually present rather than to a population, (ii) retains its in-class skill, and (iii) applies to any self-play policy and pays its compute once. On Overcooked and level-based foraging, against held-out partners from the modelled family, TTBR recovers a large share of the coordination lost with strangers and beats a generalist trained on the whole partner family where responses conflict. Where nothing conflicts a generalist suffices, and outside the modelled family TTBR degrades gracefully but has no edge, while planning keeps one.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning trains a team together and evaluates the team it trained. Deployment rarely honours that arrangement. A robot joins a team built by other groups (Genter et al., 2017), a driving policy merges among human drivers (Schwarting et al., 2019), an assistive agent works alongside a person (Javdani et al., 2018), and a controller trained in simulation joins hand-written controllers already in the field (Wu et al., 2022): partners the agent never trained with. Zero-shot coordination (Hu et al., 2020) and ad hoc teamwork (Stone et al., 2010; Mirsky et al., 2022) name this problem, and across its testbeds the pattern is the same: agents that score highly with copies of themselves collapse with a stranger, because self-play converges to conventions only a self-play partner shares.

The dominant remedy is diversity at training time: fictitious co-play (Strouse et al., 2021), maximum-entropy populations (Zhao et al., 2023), trajectory diversity (Lupu et al., 2021), hidden-utility self-play (Yu et al., 2023), incompatible-policy generation (Charakorn et al., 2023), best-response diversity (Rahman et al., 2023; 2024), mixed-play conventions (Sarkar et al., 2023) and adversarial or regret-driven generation (Wang et al., 2025; Chaudhary et al., 2026) widen the training population so that the deployment partner resembles someone the agent has met, and symmetry-based methods remove arbitrary conventions (Hu et al., 2020; 2021). All of them output one policy committed before the partner is met, and that commitment has a cost no diversity can pay: when two plausible partners call for incompatible responses, one fixed policy must compromise. In one of the kitchens we study, a self-play agent loses over nine tenths of its return with a hand-written partner that parks on a single-access cell, and the response that repairs one such partner harms another. The right response depends on which partner is in the room, and that can only be learned once the interaction has begun.

Adapting after it has begun has a long history, each branch with a limit. Type-based ad hoc teamwork infers the partner's type from its behaviour and selects a response prepared in advance for it (Albrecht

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.

& Ramamoorthy, 2013; Albrecht et al., 2016; Barrett et al., 2017), so a response per type must already exist. Bayes-adaptive reinforcement learning treats the partner as hidden state (Duff, 2002; Zintgraf et al., 2020) and learns to condition on the history by meta-, in-context or open-team reinforcement learning (Duan et al., 2016; Laskin et al., 2023; Lee et al., 2023; Rahman et al., 2021; Gu et al., 2022), but the in-context agents score at or below a random agent against unseen Overcooked teammates (Jing et al., 2026). Decision-time search plans against a model of the partner (Lerer et al., 2020), paying its compute at every step and bounded by its horizon, and test-time fine-tuning has been used in self-play Hanabi (Fickinger et al., 2021), with a partner that is a copy of the agent, so nothing is identified or retained.

This paper asks: *after a short interaction with a stranger, what should test-time compute buy?* We propose *test-time best response* (TTBR). The agent observes a few episodes between its unadapted policy and the partner and forms an exact posterior over an explicit hypothesis class that pairs a grid of parametrised partner programs, covering partners no training run produced, with a library of learned policies, covering partners that resemble the agent itself. It then improves its own policy by reinforcement learning in the simulator against that posterior, anchored by self-play and a floor of library partners so that in-class skill is kept, and deploys it at no further cost. The blueprint is trained without partner data, as in zero-shot coordination, but scored after an observation and an adaptation phase: the few-shot coordination of Nekoei et al. (2023) rather than first-contact play. TTBR keeps the type-based posterior but computes the response on demand, over a class of hundreds of programs for which no response library exists; since the MAP program does as well as the mixture on three of four layouts (RQ3), that on-demand response and the explicit class, not the mixture, separate it from type selection, and a response library precomputed for our fixed grid is the rival we did not run (§5; Appendix A has the wider literature).

We study TTBR on three Overcooked layouts (Carroll et al., 2019) and level-based foraging, against scripted strangers from the modelled family, held-out neural partners and an external benchmark, and contribute: (1) the method, which turns an observed interaction into an exact posterior over programs and policies and then into an improved policy that keeps its in-class skill, closing a large part of the gap to self-play on every layout, across three blueprints, on a disjoint second stranger set and against neural partners; (2) a theory of when identification matters: the regret of a posterior response is bounded by the conflict among candidate partners times the residual posterior odds (Theorem 1), so identification is worthless where responses do not conflict or the observation is uninformative, and both predictions hold, identification paying exactly on the two layouts where conflict is measured; (3) a matched-information comparison of the two uses of one posterior, in which decision-time planning recovers most but not all of TTBR's gain where responses conflict and all of it where they do not, and the two compose; (4) a controlled account of where the method does not help: outside the hypothesis family, with partners whose observation is uninformative, and where no responses conflict.

2 Preliminaries

Decentralised POMDPs. We work in two-player common-payoff decentralised POMDPs (Oliehoek & Amato, 2016), tuples $\langle \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, P, R, \Omega^1, \Omega^2, Z, T \rangle$ of a state space, action and observation sets per agent $i \in \{1, 2\}$, a transition kernel $P(s' \mid s, a^1, a^2)$, a shared reward $R(s, a^1, a^2)$, an observation function $Z(o^1, o^2 \mid s')$ and a horizon T. Agent i acts by a policy $\pi^i(a^i_t \mid h^i_t)$ on its private history of observations and actions, and the return $\sum_{t \le T} R(s_t, a^1_t, a^2_t)$ is common to both: fully cooperative but decentralised, since neither agent sees the other's observations or policy.

Cooperative games with an unknown partner. We control the first agent, the *ego*, whose policy π maps its history to a distribution over actions. The second, the *partner*, follows a policy τ we do not control. Write $V(\pi, \tau)$ for the expected episodic return of the pair, $V^\star_\tau = \max_\pi V(\pi, \tau)$ for the value of a best response $\mathrm{BR}(\tau)$, and $\mathrm{Reg}_\tau(\pi) = V^\star_\tau - V(\pi, \tau)$ for the regret of π against τ. The deployed agent starts from a *blueprint* π_0 trained in self-play. For a partner τ its *coordination gap* is $V(\pi_0, \pi_0) - V(\pi_0, \tau)$, the quantity every method in this paper tries to recover.

Partner types and the posterior. Uncertainty about the partner is a finite hypothesis class $\mathcal{H} = \{\tau_1, \ldots, \tau_n\}$ with prior p, and the partner is τ_θ for an index θ drawn from p. After N observed

partner actions $a_1, \ldots, a_N$ at partner histories $h_1, \ldots, h_N$ the posterior is

$$b_N(i) \;\propto\; p_i \prod_{t \le N} \tau_i(a_t \mid h_t), \tag{1}$$

exact whenever every $\tau_i(\cdot \mid h_t)$ can be evaluated on the recorded interaction, as it can for programs and for stored policies replayed on the recorded observations.

Protocol and assumptions. Zero-shot coordination scores an agent in cross-play with partners it never trained with, and ad hoc teamwork lets it adapt during the interaction. Our protocol confines that adaptation to a phase between episodes, the few-shot coordination (Nekoei et al., 2023) of a blueprint trained without partner data. In the *observation phase* the unadapted blueprint plays E episodes with the partner on each seat it can occupy, and the partner's actions and the states at which it took them are recorded. In the *adaptation phase* the agent computes for as long as its budget allows, with (A1) a simulator of the environment, as decision-time search assumes, but (A3) no further interaction with the partner. In the *deployment phase* the adapted agent is scored with the same partner. We assume (A2) that the partner is in, or behaves like a member of, $\mathcal{H}$, and (A4) that it is stationary, the same policy in all three phases. Compute is accounted in simulator steps, one axis for methods that pay once and methods that pay at every step (§5 and Appendix H return to A1–A4).

3 Method: Test-Time Best Response

§3.1 asks when the partner's identity matters at all and shows that identifying it can only help where the candidate partners' best responses conflict; §3.2 shows that identification from the observation phase is exact and cheap, and says when it is not; §3.3 spends the compute on the ego's own policy and combines the two results into a regret bound.

3.1 When does the partner's identity matter?

A response fitted to one partner may cost value with another, so whether identifying the partner can help at all is a question about how far the candidate partners' best responses disagree.

Definition 1 (conflict and value of identification)**.** For types $j, k \in [n]$ the *conflict* of j towards k is $\kappa(j \to k) = \mathrm{Reg}_{\tau_k}(\mathrm{BR}(\tau_j))$, the regret type k suffers when met with a best response to type j (for a fixed choice of best responses). For weights $\omega \in \Delta^n$ the *value of identification* is

$$\mathrm{VI}(\omega) \;=\; \textstyle\sum_k \omega_k V^\star_{\tau_k} \;-\; \max_\pi \sum_k \omega_k V(\pi, \tau_k),$$

the return of an ego that knows the type minus the best return any single policy achieves on average.

$\mathrm{VI}(\omega)$ is the expected value of perfect information about the type (Raiffa & Schlaifer, 1961; Howard, 1966), and κ is the cross-best-response regret from which minimum coverage sets are built (Rahman et al., 2024): a family needs more than one response exactly when $\kappa > 0$ somewhere.

Proposition 1 (a single response for a mixture)**.** *Let $\pi_\omega \in \arg\max_\pi \sum_k \omega_k V(\pi, \tau_k)$. Then for every j with $\omega_j > 0$,*

$$\mathrm{Reg}_{\tau_j}(\pi_\omega) \;\le\; \frac{1}{\omega_j} \sum_{k \neq j} \omega_k \, \kappa(j \to k), \qquad \textit{and} \qquad \mathrm{VI}(\omega) \;\le\; \min_j \sum_{k \neq j} \omega_k \, \kappa(j \to k).$$

Corollary 1 (no conflict, no need to identify)**.** *If there is a selection of best responses with $\kappa(j \to k) = 0$ for all j, k in the support of ω, then π_ω is a best response to every type in the support and $\mathrm{VI}(\omega) = 0$: no observation of the partner, however informative, can improve on the best single policy for the family.*

Proofs are in Appendix C. The value of knowing the partner is bounded by the conflict among the partners one might meet: where a family admits a common best response, one policy trained against the whole family is the right agent. Whether the value is realised where conflict exists is empirical (RQ1 measures κ).

3.2 IDENTIFICATION FROM OBSERVATION

A library of policies produced by training covers only partners some training run produced, so we build $\mathcal{H} = \mathcal{P} \cup \mathcal{L}$ from two kinds of hypothesis. $\mathcal{P}$ is a grid over the knobs of a family of parametrised partner *programs*, hand-written heuristics with structural and noise knobs (Appendix B); a program is a closed-form map from the simulator state and seat to an action distribution, so its likelihood in (1) is exact and a grid of hundreds costs nothing to train. $\mathcal{L}$ is the library of self-play checkpoints from the blueprint's own training, covering partners that behave like learned agents. The prior puts half its mass on each class, uniformly within it; the posterior is filtered exactly along each recorded episode and averaged over the E episodes. How much an interaction can identify is a quantity:

Lemma 1 (posterior odds after observation)**.** *Let the partner be $\tau_\theta \in \mathcal{H}$ and let $a_1, \dots, a_N$ be its actions at histories $h_1, \dots, h_N$ during an interaction with any ego policy. For a rival $i \neq \theta$ let $\rho_t(i) = \sum_a \sqrt{\tau_\theta(a \mid h_t)\, \tau_i(a \mid h_t)}$ be the Bhattacharyya coefficient of the two predictions at h_t and $D_N(i) = -\sum_{t \le N} \log \rho_t(i)$ the* evidence *accumulated against i. Then for every $d \ge 0$,*

$$\mathbb{E}_\theta\Big[\sqrt{b_N(i)/b_N(\theta)}\; \mathbf{1}\{D_N(i) \ge d\}\Big] \;\le\; \sqrt{p_i/p_\theta}\; e^{-d}.$$

Consequently, for the odds against the truth $O_N = (1 - b_N(\theta))/b_N(\theta)$ and the event $\mathcal{E}_d = \{D_N(i) \ge d \text{ for all } i \neq \theta\}$, $\mathbb{E}_\theta[\sqrt{O_N}\, \mathbf{1}\{\mathcal{E}_d\}] \le S_\theta e^{-d}$ with $S_\theta = \sum_{i \neq \theta} \sqrt{p_i/p_\theta}$.

Nothing is assumed about the environment or the ego's play. $D_N(i)$ measures how much the partner's observed behaviour disagreed with what rival i would have done at the states actually visited: one visited state at which two deterministic programs choose different actions eliminates the rival, while a rival that prescribes the partner's own behaviour at every visited state leaves $D_N(i) = 0$ however long the observation lasts. With a uniform prior over n programs $S_\theta = n - 1$, so evidence of order $\log n$ nats (under six here) suffices: one decisive disagreement, or about a hundred steps against a rival whose predictions overlap the partner's by 0.95 per step, and Appendix F finds the most likely program fixed within 25 to 50 steps of an episode. The posterior separates only what the blueprint's play elicited, a fact RQ4 and Appendix G will need.

3.3 SPENDING THE COMPUTE ON THE POLICY

What the blueprint lacks against a stranger is a best response: with a partner that parks on a cell or never fetches a plate, the blueprint must perform a half of the task it never learned. A best response to a posterior is a policy, and we compute it as policies are computed, by reinforcement learning in the simulator from the blueprint's parameters against partners drawn from the belief. Fine-tuning against the belief alone forgets in-class play (the worst held-out self-play partner loses 56 points, and a larger self-play share still leaves -28, a KL trust region -19), whereas a uniform floor of library partners keeps every such partner above the blueprint ($+16.6$ to $+30.7$) at 1.02 of the stranger gain, and floor weights 0.1, 0.25, 0.5 trade stranger gains $+70.0$, $+69.7$, $+55.9$ against in-class gains $+12.8$, $+24.0$, $+33.4$ (Appendix F). The adaptation phase therefore maximises, for U updates of PPO (Schulman et al., 2017),

$$J(\pi) \;=\; \alpha_0\, V(\pi, \pi_0) + \alpha_L \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} V(\pi, \tau_\ell) + \alpha_b \sum_i b_N(i)\, V(\pi, \tau_i), \quad (\alpha_0, \alpha_L, \alpha_b) = (\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{2}), \tag{2}$$

where each parallel environment draws its partner kind from $(\alpha_0, \alpha_L, \alpha_b)$ and then its partner: the blueprint itself, a library member uniformly, or a hypothesis from the posterior. The posterior term is the best response the observation asks for, and the self-play term and the uniform *library floor* keep the skills the agent already has, at a price in stranger regret that Corollary 2 (Appendix C) bounds by the same conflict quantity. For the posterior term itself, Proposition 1 and Lemma 1 combine into a guarantee.

Theorem 1 (regret of the posterior response)**.** *Let $\pi_b \in \arg\max_\pi \sum_i b_N(i)\, V(\pi, \tau_i)$ be a best response to the posterior after the observation phase, and let $\bar\kappa_\theta = \max_{k \neq \theta} \kappa(\theta \to k)$ be the largest conflict of the true type towards a rival. Then (i) $\mathrm{Reg}_{\tau_\theta}(\pi_b) \le b_N(\theta)^{-1} \sum_{k \neq \theta} b_N(k)\, \kappa(\theta \to k) \le \bar\kappa_\theta\, O_N$, and (ii) for every $d \ge 0$ and $\eta \in (0, 1]$, $\mathbb{P}_\theta\big(\mathcal{E}_d \text{ and } \mathrm{Reg}_{\tau_\theta}(\pi_b) > \bar\kappa_\theta\, S_\theta^2\, e^{-2d}/\eta^2\big) \le \eta$.*

The regret of the posterior response is the product of two factors, the conflict among the partners that could have been met and the odds the observation left against the truth, and three regimes follow.

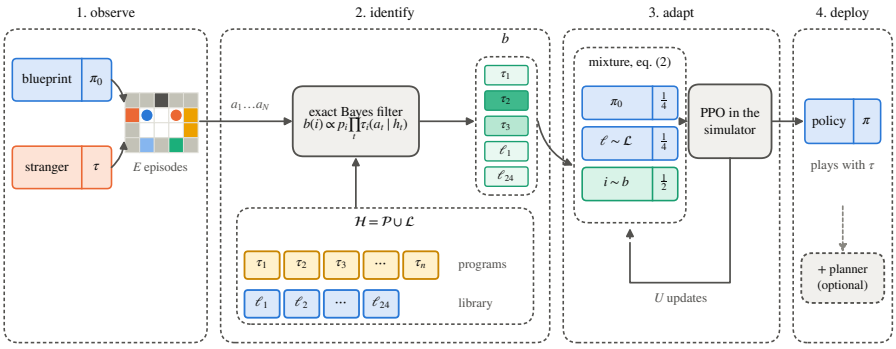

Figure 1: **Test-time best response.** 1) The blueprint π_0 plays E episodes with the stranger τ. 2) An exact Bayesian filter over the hypothesis class $\mathcal{H}$ turns the recorded actions into the belief b. 3) PPO in the simulator improves π_0 against the mixture of itself, the library floor and b. 4) The improved policy is deployed, optionally with a planner on top.

Where nothing conflicts ($\bar{\kappa}_\theta = 0$) the posterior is irrelevant and the ego might as well be a generalist. Where the observation eliminates every rival ($D_N = \infty$) the response is a best response to the partner itself. Where the blueprint's play elicits no distinguishing behaviour ($D_N \approx 0$) the odds keep their prior value and identification buys nothing, whatever compute is spent afterwards. The conflict-weighted form of (i) covers ties: a rival that behaves as the true type does on the observed seat has $\kappa(\theta \to k) = 0$ and may keep any mass at no cost (the exact ties of Appendix F). In (ii) the factor $1/\eta^2$ is the price of Markov's inequality on the square-root odds, a constant that $\log(1/\eta)$ extra nats of evidence absorb. The theorem's object is the exact best response to the posterior; the deployed policy maximises (2) only approximately, and Proposition 2 carries Corollary 2 to an ε-approximate maximiser at an additive $\varepsilon/(\alpha_b b_N(\theta))$ on top of the anchors' price, a floor identification cannot remove (Appendix C also covers a partner outside the class). Figure 1 summarises the method (Algorithm 1 in Appendix D, with a post-hoc safeguard that keeps the blueprint if the simulated return against the belief never rose).

4 Experiments

4.1 Environments, partners and blueprints

Layouts. We use three OvercookedV2 layouts (Carroll et al., 2019; Rutherford et al., 2024; Gessler et al., 2025) and a level-based foraging grid (Albrecht & Ramamoorthy, 2013; Papoudakis et al., 2021b; Bonnet et al., 2024) (schematics in Figure 6, Appendix B), with episodes of 400 and 100 steps. *Cramped room* has cells only one agent can pass, so where the partner stands decides what the ego can reach; *test-time wide* separates a cook seat from a signaller seat whose convention over a one-step indicator the partner may or may not follow; *asymmetric advantages* gives the seats different reach and changing recipes; *LBF* needs the agents to pool their levels.

Partners. Three kinds of partner are held out from every training run. *Strangers* are hand-written heuristic partners from the same parametrised family as $\mathcal{P}$; apart from four noise-free defaults their knob values are off the hypothesis grid (Appendix B), so identification is of approximate parameters, never of a copy. There are 14 on cramped room, where the *parkers* among them occupy a single-access cell for most of the episode, and 12 on each other layout, and a *second, disjoint set* of six per Overcooked layout, never touched by TTBR before, serves only to replicate the main result. *Siblings* are eight independently trained self-play seeds, the neural partners of the usual cross-play test. The 20 heuristic teammates of the ICRL4AHT benchmark (Jing et al., 2026) on wide are partners from outside our family, scored under the benchmark's protocol.

Blueprints, library and hypothesis class. Blueprints are self-play PPO agents trained for 3×10^7 steps, three seeds, each reported. The library $\mathcal{L}$ holds 24 checkpoints of eight self-play seeds, and

$\mathcal{P}$ is the layout's knob grid of 72 to 288 programs. The observation phase plays $E = 16$ episodes per seat (32 on LBF, whose episodes are shorter), and the adaptation phase runs $U = 100$ PPO updates (6.6×10^6 simulator steps, about 70 minutes on eight CPU cores) with the weights of (2) (Appendix D).

4.2 Baselines, instruments and statistics

Train-time baselines. We compare against self-play (SP), fictitious co-play (Strouse et al., 2021, FCP), maximum-entropy population training (Zhao et al., 2023, MEP) with both bonus weights (the better reported) and five further population variants (Appendix B), all with five ego seeds, and against a baseline no published method reaches: the *program-generalist*, trained directly against the whole program family. Every population member and every ego, the generalist included, trains for the blueprint's budget of 3×10^7 steps, and our MEP trains its ego with FCP's uniform sampling rather than MEP's prioritised sampling. The best case of any partner-generation method for these strangers is to produce this very family, which the generalist is handed, so its results stand for what partner generation could achieve here under a uniform curriculum at this budget (a prioritised or adversarial curriculum over the same family could still do better).

Test-time comparisons at matched information. A posterior can be consumed in two ways. *Planning* uses it at decision time: at every real step each candidate action is scored by rolling the blueprint forward for H steps against partners drawn from b and the best is taken (Lerer et al., 2020), so there is no adaptation phase and the cost is paid at every step. *Policy improvement* uses it once, for U updates, after which the improved policy acts at forward-pass cost. Both approximate the π_b of Theorem 1 and differ in what limits them, the horizon H or the budget U. We hold everything but the use of the posterior fixed: the planner rolls out the same blueprint against partners from the same class $\mathcal{H}$ and, in the arm that matters, from TTBR's own posterior after the same observation phase, at $H \in \{8, 16, 32, 64\}$, and three ablated planners bracket it (Appendix D). On the improvement side, the *true-program* fine-tune, our *oracle best response*, is the ceiling for identification, and a best response to the *uniform prior* over $\mathcal{P}$ is one policy per layout with no observation.

Instruments and statistics. Policies are scored with 32 episodes per partner per seat under a fixed evaluation seed, so every gain is a paired difference against the blueprint with the same partner. Means are over strangers, intervals are 95% t-intervals over strangers or over blueprints $\times$ strangers where pooled, and comparisons with planning use the 12 strangers common to the Overcooked layouts (Appendix D). Some strangers behave identically on a layout (10 distinct behaviours among the 14 of cramped room, 11 of 12 on wide, 9 of 12 on asymmetric advantages), so Table 2 also gives intervals over distinct behaviours. Decision rules were fixed before the jobs ran (Appendix I), unregistered analyses are marked *post hoc*, and no multiplicity correction is applied: the registered rules fix which comparisons are confirmatory. One instrument defect must be disclosed: until the fourth day of the study the search evaluator fed every trained network zeros on one input channel, the partner's previous action, which a self-play reference scoring -63 with a copy of itself exposed. Every policy was re-scored with the repaired instruments, no number measured before the repair is quoted, and every pre-registered verdict taken before it survived (Appendix I).

4.3 Results

RQ1: Is the coordination gap a conflict gap that train-time diversity cannot close? Figure 2 scores every train-time method against the strangers on cramped room and wide, the two layouts where partner responses turn out to conflict (numbers in Table 3, Appendix E). On cramped room the blueprint scores 220 with a copy of itself and 121 with the strangers, and FCP, MEP and the five population variants all score below it, between 58 and 76, not for lack of training: the same egos score 90 to 99 with the eight noisy held-out strangers but far less with the other six, the four noise-free defaults and the two parkers of the deliver-self pair (FCP 0 to 60, MEP below 14). On wide every train-time method is negative and only the program-generalist, trained on the strangers' own family, scores positively, while the oracle best response (a fine-tune to each stranger's true program, the ceiling for identification) reaches 195 on cramped room and 81 on wide. Conflict says why: we fine-tuned a best response to every stranger's true program and scored it against every other stranger of its layout, recording the *harm* $h(j \to k) = V(\pi_0, \tau_k) - V(\mathrm{BR}(\tau_j), \tau_k)$, how much worse the

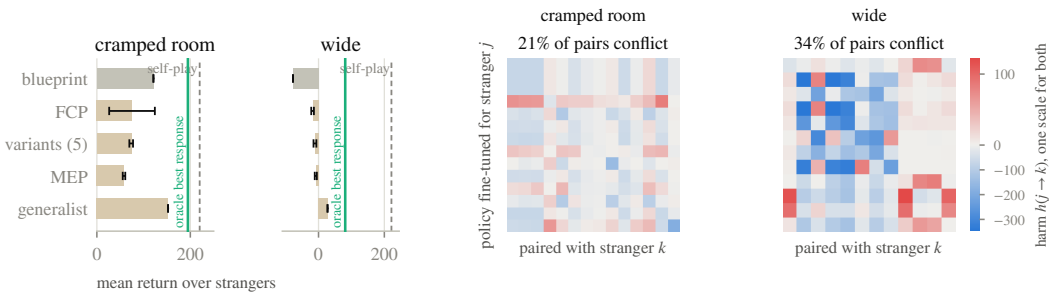

Figure 2: **Train-time diversity does not close the gap, and conflict is why.** Left: mean return over strangers for the blueprint and the train-time methods, with the self-play level and the oracle best response, a policy fine-tuned to each stranger's true program. Right: harm $h(j \to k)$, how much worse the policy fine-tuned for stranger j (row) does with stranger k (column) than the unadapted blueprint; red is harm, blue is gain, one colour scale for both panels (labels in Figure 7).

response to j does with k than not adapting (so $h \leq \kappa$). On cramped room 21% of same-seat pairs have $h > 10$ and on wide 34%, while on asymmetric advantages and LBF none do; accordingly, the best single fine-tune recovers only about half of what the per-partner fine-tunes achieve on the first two layouts but nearly all of it on the last two, and the generalist, which must hold one response for the whole family, ends below the blueprint on the parkers whose routes conflict. So on cramped room and wide the gap is a conflict gap that train-time diversity cannot close; the other two layouts are the no-conflict control Corollary 1 asks for (RQ3).

RQ2: How much of the gap does TTBR recover, and does it hold up? Figure 3 gives the main result stranger by stranger, against the generalist and the oracle best response, and Table 2 (Appendix E) the summary for four layouts and three blueprints (per stranger in Figure 8). Pooled over the blueprints, the mean gain is $+99.4$ $[+79.8, +119.0]$ on cramped room, about 0.7 of the gap to self-play. On wide, where the blueprint's convention fails with most strangers, it is $+158.8$ $[+106.7, +210.9]$. On asymmetric advantages it is $+47.1$ $[+38.4, +55.8]$, and on LBF $+0.17$ $[+0.14, +0.20]$ on a unit scale. The registered replication rule for the other two blueprints (at least half of the first blueprint's gain) holds in every cell but one: on asymmetric advantages blueprint 1002 gains $+29.4$ against a threshold of 35.7 and reads NOT REPLICATED. The second, disjoint set of six strangers per layout replicates the result on cramped room and asymmetric advantages (6/6 positive, intervals above zero) and is directionally positive but INCONCLUSIVE under the registered rule on wide (4/6, $+179.7$ $[+12.5, +346.9]$), with the same two cook-seat losses (Appendix G). Retention against four held-out siblings holds under the registered floor of -10 (3.5 to 4.5% of the Overcooked self-play returns), which on cramped room binds every fine-tune's mean change over the four siblings and on wide the mean over fine-tunes: the first blueprint gains with every fine-tune (worst $+16.1$, $+16.9$, $+19.0$), and the one fine-tune below -10 anywhere, blueprint 1002 on wide at -17.1, sits under the wide rule's mean of -6.1, which passes (Table 2 lists every worst value). The floor that secures retention costs no stranger gain (Appendix F). All numbers are ungated, the conservative choice: the deployment check of Algorithm 1 would move the wide mean to $+172.4$ and nothing else.

RQ3: When is test-time identification needed? Figure 4 holds the method fixed and varies only what the mixture's partner term knows: nothing (a best response to the uniform prior over $\mathcal{P}$), the family (the converged generalist), the most likely program (MAP), the posterior (TTBR) or the partner itself (true program). On cramped room the posterior beats the prior by $+83.6$ $[+51.8, +115.4]$ and the generalist by $+35.8$ $[+14.0, +57.6]$, and on wide by $+70.4$ and $+48.4$. On asymmetric advantages the prior suffices and the generalist is *better* than TTBR by 11.4 [3.3, 19.5], and on LBF the prior and posterior policies are indistinguishable. Identification pays exactly on the two layouts where RQ1 found conflict, as Corollary 1 and Theorem 1 predict, and it costs little: the true program improves on the posterior by at most $+6.6$. Nor is the mixture the main ingredient: fine-tuning against the MAP program alone does as well on cramped room, asymmetric advantages and LBF, and on wide the mixture is ahead by $+12.8$ $[-9.6, +35.3]$ (registered verdict MAP SUFFICES on

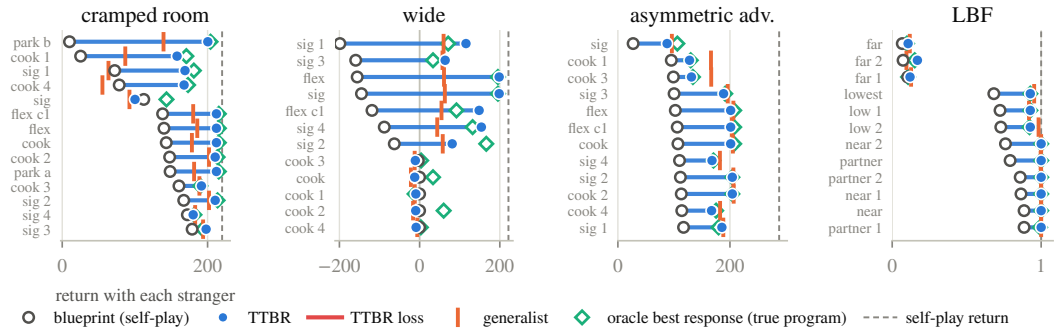

Figure 3: **Main result, stranger by stranger** (first blueprint), rows sorted by the blueprint's return: hollow circle, blueprint; filled circle, TTBR (red where it loses); orange tick, generalist; hollow diamond, oracle best response.

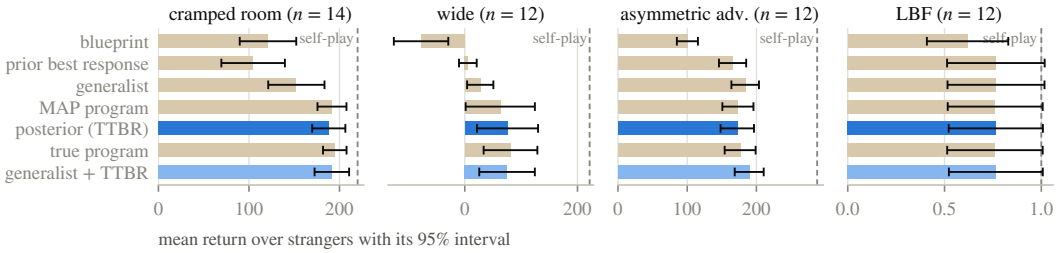

Figure 4: **The identification ladder.** Mean return over strangers with its 95% interval as the mixture's partner term knows nothing, the family, the MAP program, the posterior or the true program, and for TTBR started from the generalist. Dashed: self-play.

three layouts, INCONCLUSIVE on wide, Appendix F), so identification buys mostly the right partner to practise against, not belief averaging. Started from the converged generalist, TTBR adds $+38.3$ $[+16.4, +60.2]$ on cramped room and $+42.2$ $[+11.3, +73.2]$ on wide, ending within four points of TTBR from self-play, and adds $+6.6$ on asymmetric advantages and nothing on LBF, as Corollary 1 predicts.

RQ4: Given the posterior, plan or improve? Figure 5 puts every test-time arm on one axis of simulator steps, with the generalist's zero-cost level for reference. The comparison is fair on two counts (Appendix F): the planner is used at its best (its gain is limited by the horizon, not the rollouts, and peaks around $H = 32$, so we compare at $H = 32$ and, on the first blueprint, 64), and at matched information (handed TTBR's posterior, the planner gains as much as with the partner's exact program, so only the use of the posterior differs). (1) *From the same posterior, policy improvement extracts more where responses conflict.* Pooled over three blueprints at $H = 32$, TTBR is ahead by $+26.4$ $[+16.8, +36.0]$ on cramped room and $+75.2$ $[+50.0, +100.4]$ on wide, and at $H = 64$ on the first blueprint still by $+12.9$ $[+7.0, +18.8]$ on all 12 cramped-room strangers, while on wide the interval includes zero (far ahead on the signaller seat, behind on the cook seat it fails on); on asymmetric advantages the two tie. (2) *The two fail on different partners, and Theorem 1 says why.* Improvement fails where the observation carried no evidence: with the five cook-seat strangers of wide, the zero-evidence case of the bound, TTBR loses about ten points each while the planner, whose lookahead sees the partner respond within the episode, gains up to $+40$ at $H = 64$. Planning falls short where the right response pays off only over the whole episode: with the three parking strangers of cramped room it trails TTBR by 43 to 70 points at $H = 32$, and with the seven signaller-seat strangers of wide, whose convention must be held all episode, TTBR gains up to $+362$ against the planner's $+239$. (3) *The two compose where improvement leaves value.* Planning on top of the TTBR policy adds $+48.9$ $[+31.1, +66.6]$ on wide, rescuing the cook seat, $+24.1$ $[+17.2, +30.9]$ on asymmetric advantages, and nothing on cramped room, where the adapted policy already scores 188 of the self-play 220. (4) *Cost.* TTBR pays 3.3 to 6.5 M simulator steps per partner, once, then acts

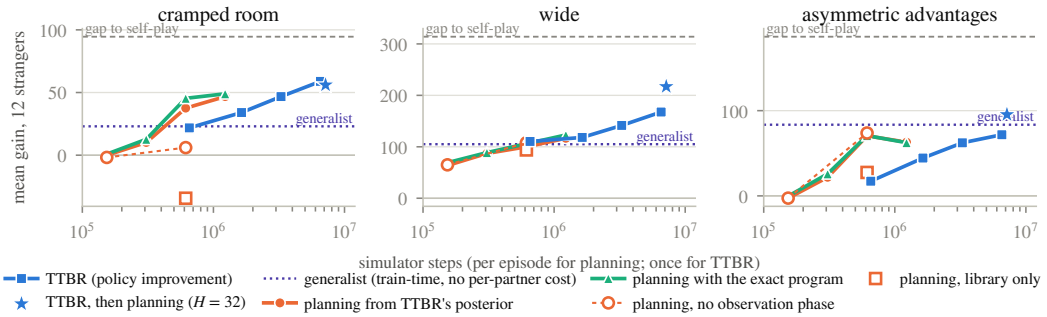

Figure 5: **Planning or policy improvement, from the same posterior.** Mean gain over the 12 common strangers against simulator steps, paid per episode by the planners and once by TTBR. Dotted: the generalist, which costs nothing per partner.

at forward-pass cost, whereas planning pays 0.6 M steps per episode at $H = 32$ with no adaptation phase, so against planning TTBR is repaid within 3 to 11 episodes; against a train-time method it is repaid only in return, the generalist paying nothing per partner and scoring 36 and 48 points lower where responses conflict (RQ3). Where responses conflict, the compute is better spent on improving the policy, with planning on top where the observation left the partner unidentified; where nothing conflicts, either use suffices.

RQ5: Does it extend beyond scripted partners? Two tests leave the scripted family (Figure 9 and Table 4, Appendix E). Against the eight held-out self-play seeds TTBR gains $+46.6$ $[+14.2, +79.1]$ on cramped room, $+44.0$ $[+23.9, +64.0]$ on asymmetric advantages and $+76.8$ on wide with a wide spread; the gain transfers to the seven seeds the ego never observed, so most of it is generic cross-play skill rather than identification, and planning with the same library as its partner model falls below TTBR at $H = 32$ on all three layouts. The 20 ICRL4AHT teammates lie outside our family and invert the picture: under the benchmark's protocol both test-time methods beat the published in-context baselines, search clearly ($+17.0$) and TTBR barely ($+2.7$ $[+1.5, +4.0]$), indistinguishable from a generalist ($+2.7$) and a best response to the prior ($+2.4$). Outside the class Theorem 1 is silent: identification does no harm there, but has no edge.

5 Limitations and Conclusion

A2 is the load-bearing assumption: our strangers come from the same parametrised family as $\mathcal{P}$, off its grid but inside its span, and outside the family TTBR does no harm but has no edge (RQ5). Two natural rivals were not run: a library of best responses precomputed for the fixed grid and selected by the same posterior, a serious contender in-family given the MAP result of RQ3 (so RQ4 shows improvement beats planning, not selection), and an observation phase shorter than $E = 16$ episodes, which Appendix F suggests would suffice. A4 was never tested: every partner is a fixed policy throughout, and RQ1's worst harm pairs (-83, -121) are the exposure of an adapted policy to a partner that changes after adaptation. The strangers hold 10 of 14, 11 of 12, 9 of 12 and 12 of 12 distinct behaviours on the four layouts, so intervals over strangers are optimistic; two intervals on wide include zero; the asymmetric-advantages intervals of blueprints 1001 and 1002 are post hoc. We do not claim that improvement beats planning in general, that planning is brittle under posterior uncertainty, that the posterior mixture matters, or that identification is needed everywhere (Appendix H).

We introduced test-time best response (TTBR): after a short observation phase, an exact posterior over an explicit class of programs and policies, and reinforcement learning against it, anchored to keep in-class skill. We bounded the regret of the posterior response by conflict times residual odds, extended it to the deployed anchored, approximate policy, and compared improvement and planning at matched information. Against strangers from the modelled family, TTBR recovers much of the lost coordination, beats a generalist where responses conflict, composes with planning where value remains, and degrades gracefully outside the family.

Reproducibility statement

Appendix D gives the network, the PPO settings, the observation and adaptation phases, the evaluation protocol and the compute of every run. Appendix B lists the layouts, the partner programs with their knobs, every stranger and the budget of every baseline. Appendix C contains the proofs. Appendix E reports the per-stranger values behind every number in the paper, Appendix I the decision rules, their verdicts, and the evaluator defect and its repair, Appendix J the lines that were closed before the method existed, and Appendix K the use of language models. The code, the configuration files of every run, the dated decision-rule files with their commit timestamps, and the script that generates every figure and table from the stored results will be released.

AI use statement

In this work, we used generative AI tools as coding and writing assistants: to implement and test the experiment code under the authors' specifications, to run and monitor the cluster jobs, to draft the analysis scripts that read the pre-registered rules, and to draft and edit the text. Generated code was verified and tested for correctness, and the text was reviewed by the authors, who take responsibility for it. We have not used generative AI tools to generate research ideas, decision rules or experimental designs; every rule was written by the authors before the corresponding jobs ran (Appendix K).

References

Ekin Akyürek, Mehul Damani, Linlu Qiu, Han Guo, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for abstract reasoning. *arXiv preprint arXiv:2411.07279*, 2024.

Stefano V. Albrecht and Subramanian Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258, 2018.

Stefano V. Albrecht, Jacob W. Crandall, and Subramanian Ramamoorthy. Belief and truth in hypothesised behaviours. *Artificial Intelligence*, 235:63–94, 2016.

Samuel Barrett, Avi Rosenfeld, Sarit Kraus, and Peter Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242, 2017.

Robert H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 37(1), 1966.

Dimitri P. Bertsekas. *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. Athena Scientific, 2020.

Clément Bonnet, Daniel Luo, Donal Byrne, Shikha Surana, Sasha Abramowitz, Paul Duckworth, Vincent Coyette, Laurence I. Midgley, Elshadai Tegegn, Tristan Kalloniatis, Omayma Mahjoub, Matthew Macfarlane, Andries P. Smit, Nathan Grinsztajn, Raphaël Boige, Cemlyn N. Waters, Mohamed A. Mimouni, Ulrich A. Mbou Sob, Ruan de Kock, Siddarth Singh, Daniel Furelos-Blanco, Victor Le, Arnu Pretorius, and Alexandre Laterre. Jumanji: a diverse suite of scalable reinforcement learning environments in JAX. In *International Conference on Learning Representations (ICLR)*, 2024.

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit 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.

Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. Generating diverse cooperative agents by learning incompatible policies. In *International Conference on Learning Representations (ICLR)*, 2023.

Paresh Chaudhary, Yancheng Liang, Daphne Chen, Simon S. Du, and Natasha Jaques. Improving human-ai coordination through online adversarial training and generative models. In *International Conference on Learning Representations (ICLR)*, 2026.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst, 2002.

Arnaud Fickinger, Hengyuan Hu, Brandon Amos, Stuart Russell, and Noam Brown. Scalable online planning via reinforcement learning fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Katie Genter, Tim Laue, and Peter Stone. Three years of the RoboCup standard platform league drop-in player competition: Creating and maintaining a large scale ad hoc teamwork robotics competition. *Autonomous Agents and Multi-Agent Systems*, 31(4):790–820, 2017.

Tobias Gessler, Benjamin Ellis, Tin Dizdarevic, Andrei Lupu, Anisoara Calinescu, and Jakob N. Foerster. OvercookedV2: Rethinking overcooked for zero-shot coordination. In *International Conference on Learning Representations (ICLR)*, 2025.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 8(5–6), 2015.

Pengjie Gu, Mengchen Zhao, Jianye Hao, and Bo An. Online ad hoc teamwork under partial observability. In *International Conference on Learning Representations (ICLR)*, 2022.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning (ICML)*, 2020.

Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning (ICML)*, 2021.

Shervin Javdani, Henny Admoni, Stefania Pellegrinelli, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared autonomy via hindsight optimization for teleoperation and teaming. *The International Journal of Robotics Research*, 37(7):717–742, 2018.

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.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations (ICLR)*, 2023.

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Adam Lerer, Hengyuan Hu, Jakob Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *AAAI Conference on Artificial Intelligence*, 2020.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning (ICML)*, 2021.

Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. A survey of ad hoc teamwork research. In *European Conference on Multi-Agent Systems (EUMAS)*, 2022.

Hadi Nekoei, Xutong Zhao, Janarthanan Rajendran, Miao Liu, and Sarath Chandar. Towards few-shot coordination: Revisiting ad-hoc teamplay challenge in the game of Hanabi. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2023.

Andrew Ni, Simon Stepputtis, Stefanos Nikolaidis, Michael Lewis, Katia P. Sycara, and Woojun Kim. Theory of mind guided strategy adaptation for zero-shot coordination. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026.

Rupal Nigam, Niket Parikh, Hamid Osooli, Mikihisa Yuasa, Jacob Heglund, and Huy T. Tran. Zero-shot coordination in ad hoc teams with generalized policy improvement and difference rewards. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2026.

Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

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)*, 2021a.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021b.

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International Conference on Machine Learning (ICML)*, 2018.

Arrasy Rahman, Niklas Höpner, Filippos Christianos, and Stefano V. Albrecht. Towards open ad hoc teamwork using graph-based policy learning. In *International Conference on Machine Learning (ICML)*, 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, 1961.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktäschel, Chris Lu, and Jakob Nicolaus Foerster. JaxMARL: Multi-agent RL environments and algorithms in JAX. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. Diverse conventions for human-ai collaboration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(50): 24972–24978, 2019.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI Conference on Artificial Intelligence*, 2010.

DJ Strouse, Kevin R. McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018.

Caroline Wang, Arrasy Rahman, Benjamin Nativi, Johnny Liu, Jiaxun Cui, Yoonchang Sung, and Peter Stone. ROTATE: Regret-driven open-ended training for ad hoc teamwork. *arXiv preprint arXiv:2505.23686*, 2025.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Xihuai Wang, Shao Zhang, Wenhao Zhang, Wentao Dong, Jingxiao Chen, Ying Wen, and Weinan Zhang. ZSC-Eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Cathy Wu, Abdul Rahman Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M. Bayen. Flow: A modular learning framework for mixed autonomy traffic. *IEEE Transactions on Robotics*, 38(2):1270–1286, 2022.

Sarah A. Wu, Rose E. Wang, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.

Chao Yu, Jiaxuan Gao, Weilin Liu, Botian Xu, Hao Tang, Jiaqi Yang, Yu Wang, and Yi Wu. Learning zero-shot cooperation with humans, assuming humans are biased. In *International Conference on Learning Representations (ICLR)*, 2023.

Jaleh Zand, Jack Parker-Holder, and Stephen J. Roberts. On-the-fly strategy adaptation for ad-hoc agent coordination. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2022.

Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot human-ai coordination. In *AAAI Conference on Artificial Intelligence*, 2023.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A very good method for bayes-adaptive deep RL via meta-learning. In *International Conference on Learning Representations (ICLR)*, 2020.

A Related work

Zero-shot coordination and train-time diversity. Zero-shot coordination was posed by Hu et al. (2020) as the problem of coordinating with independently trained agents, and its first answers removed the arbitrariness of self-play conventions by exploiting symmetries of the game, in other-play and off-belief learning (Hu et al., 2020; 2021). The larger family answers it with population training: fictitious co-play trains a best response to a population of past self-play checkpoints (Strouse et al., 2021). Maximum-entropy population training adds an entropy bonus on the population's average policy so

that its members differ (Zhao et al., 2023). Trajectory diversity (Lupu et al., 2021), hidden-utility self-play (Yu et al., 2023), incompatible-policy generation (Charakorn et al., 2023), best-response diversity and minimum coverage sets (Rahman et al., 2023; 2024) and mixed-play conventions (Sarkar et al., 2023) each define what a useful population is and generate one. Open-ended and adversarial generation (Wang et al., 2025; Chaudhary et al., 2026) lets the population grow against the learner's weaknesses. Wang et al. (2024) propose best-response-normalised evaluation for these methods. All of them produce a single fixed policy, and Nekoei et al. (2023) argue that first-contact play is too strict a test of it in Hanabi and propose few-shot coordination, a handful of episodes with the partner before scoring, which is the setting of our protocol. Our train-time baselines are self-play, FCP and MEP (§4.2), and the converged program-generalist stands in for the rest by being trained on the very family the test partners come from, which is the ideal outcome of any generation method for these partners under a uniform curriculum (§4.2), and Corollary 1 says when that ideal is enough. The conflict quantity κ of Definition 1 is the cross-best-response regret from which Rahman et al. (2024) build minimum coverage sets: a family needs a coverage set larger than one exactly when $\kappa > 0$ somewhere, and Proposition 1 says how much the extra members are worth.

Type-based ad hoc teamwork and opponent modelling. Ad hoc teamwork (Stone et al., 2010; Mirsky et al., 2022) allows the agent to adapt during the interaction. The type-based line is closest to ours: HBA (Albrecht & Ramamoorthy, 2013) maintains a posterior over a set of hypothesised partner types and plays a best response to the belief, Albrecht et al. (2016) characterise when that posterior converges and what the response to it is worth when the true type is in the class and when it is not, and PLASTIC (Barrett et al., 2017) learns or selects among policies for previously met teammates by Bayesian updates on observed behaviour. Bayesian delegation (Wu et al., 2021) infers, in the same kind of kitchen, which sub-task the partner has taken on and plans against that inference. Zand et al. (2022) apply the same idea to deep population agents, inferring a posterior over a trained population and acting with a belief-weighted response. More recent variants select among an ensemble of policies by an inferred theory-of-mind embedding (Ni et al., 2026) or combine pretrained policies by generalised policy improvement (Nigam et al., 2026). All of these select or combine responses that exist before the partner is met. TTBR shares the exact Bayesian filtering and differs in what it does with the belief: it computes the response for the belief it meets, over a class of hundreds of programs for which no library of responses exists, and it uses the same belief to drive a planner. In our data the posterior mixture adds little over its most likely member: fine-tuning against the MAP program alone does as well on three of four layouts, and on wide the mixture is ahead by 12.8 with an interval including zero (§4.3), so the difference from type selection is mainly the on-demand response and the explicit program class, not the mixture. Learned opponent models (Rabinowitz et al., 2018; Papoudakis et al., 2021a) and the broader survey of agent modelling (Albrecht & Stone, 2018) address the inference side without a test-time policy-improvement step. On the theory side, Albrecht et al. (2016) prove that the posterior concentrates on types behaviourally indistinguishable from the truth and that the response to it is then optimal; Theorem 1 adds a finite-observation regret bound whose two factors, the conflict κ and the residual odds, are measured rather than assumed, and Corollary 3 its degradation under an imperfect class. A library of best responses precomputed for our fixed grid and selected by the same posterior, the type-based method in its strongest form, was not run (Appendix H).

Bayes-adaptive and meta-reinforcement learning. Treating the unknown partner as hidden state makes the problem a Bayes-adaptive POMDP (Duff, 2002; Ghavamzadeh et al., 2015), whose optimal policy is intractable. VariBAD (Zintgraf et al., 2020) and the meta-RL line from RL^2 (Duan et al., 2016) learn approximations that condition on the history, and in-context reinforcement learning distils adaptation into a sequence model (Laskin et al., 2023; Lee et al., 2023). In ad hoc teamwork, graph-based policy learning (Rahman et al., 2021) and ODITS (Gu et al., 2022) learn at train time to condition the ego on an inferred representation of its teammates, under open teams and under partial observability, from the teammates seen there. The ICRL4AHT benchmark (Jing et al., 2026) tests these on ad hoc teamwork in Overcooked and finds them at or below random against unseen heuristic teammates, recommending explicit partner inference. Our recurrent generalist, trained against the partner family, shows no sign of such conditioning on the layouts where it would pay (RQ3), and our external benchmark uses ICRL4AHT's teammates and protocol.

Decision-time search and test-time fine-tuning. SPARTA (Lerer et al., 2020) improves a blueprint policy at decision time by one-step lookahead with rollouts against a model of the partners, and is the planning arm of §4. Fickinger et al. (2021) replace the search by reinforcement-learning fine-tuning of the blueprint at test time, in Hanabi self-play, where the partner's policy is a copy of the agent's and nothing has to be identified or retained. TTBR takes the fine-tuning route against an inferred posterior over strangers, pays for retention with the library floor, and is compared with planning from the same posterior. Rollout-based improvement of a base policy is the classical setting of Bertsekas (2020).

Test-time training and test-time compute. Adapting a model on the test input has a separate history in supervised learning, from test-time training with a self-supervised loss (Sun et al., 2020) and entropy minimisation (Wang et al., 2021) to policy adaptation during deployment (Hansen et al., 2021) and, recently, test-time training for reasoning (Akyürek et al., 2024) and the study of how test-time compute should be spent (Snell et al., 2024). Our question is that last one, asked of a coordinating agent: given an observed interaction, spend the compute on planning or on improving the policy.

Programmatic hypotheses. Programmatic policies have been proposed as interpretable, verifiable alternatives to neural ones (Verma et al., 2018). We use a parametrised program family not as the ego's policy but as a hypothesis class for the partner, where its exact likelihood and zero training cost make a grid of hundreds of programs a free source of coverage that a library of learned policies cannot provide (§3.2).

B Environments, partner families and strangers

Layouts. The three Overcooked layouts are the `cramped_room_v2`, `test_time_wide` and `asymm_advantages_recipes_center` layouts of OvercookedV2 as shipped with JaxMARL 0.1.0 (Rutherford et al., 2024; Gessler et al., 2025), with 400-step episodes, negative rewards for wrong deliveries, recipes re-sampled on delivery and random initial agent positions. Cramped room and asymmetric advantages are fully observed, while wide uses an agent view of radius two and emits a one-step indicator after a correct delivery, as in ICRL4AHT (Jing et al., 2026). Level-based foraging is the 7×7 grid of Jumanji (Bonnet et al., 2024) with three foods, two agents of level at most two, forced cooperation, shared rewards and 100-step episodes, and returns lie in $[0, 1]$. Observations are the raw integer layers of each environment cast to floating point.

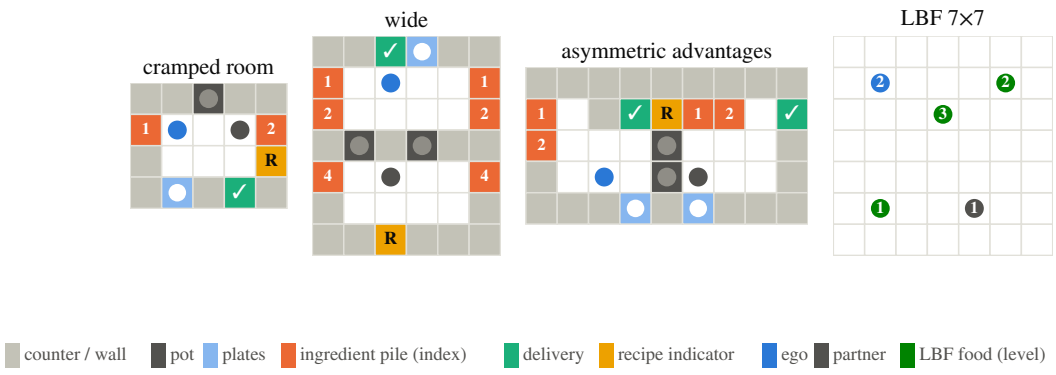

Figure 6: **The four layouts** (tile schematics, agents at their nominal start cells, ego in blue). Cramped room: one small kitchen whose pot, plates and delivery window are reached through single-access cells, so a partner that occupies one of them blocks the other agent. Wide: two rooms and two roles, a cook and a signaller who must agree on a convention over a one-step indicator, with the recipe shown on the indicator tile. Asymmetric advantages: two rooms with unequal reach and recipes that change on delivery. Level-based foraging: two agents whose levels must be pooled to collect food of a higher level.

The partner family and the program grid. Partners of the Overcooked family are heuristic programs with structural knobs `role_pref` (cook, signaller or flexible), `reads_indicator`, `signal_convention`, `plate_first`, `path_tiebreak` and `deliver_self`, an integer `signal_delay`, and noise knobs `hesitation_prob`, `wander_prob` and `wrong_action_prob`. The LBF family has `target_rule` (nearest, partner-nearest, farthest, lowest id), `path_tiebreak`, `load_rule` and the same noise knobs. The hypothesis grid $\mathcal{P}$ of a layout is the Cartesian product of the structural knobs that can change behaviour on that layout (knobs inert on a layout are left at their defaults) with the noise levels $\{0, 0.15, 0.35\}$ for hesitation and wandering and $\{0, 0.08\}$ for wrong actions: 144 programs on cramped room (`plate_first`, `path_tiebreak`, `deliver_self`), 288 on wide (`plate_first`, `path_tiebreak`, `reads_indicator`, `signal_convention`), 72 on asymmetric advantages (`plate_first`, `path_tiebreak`) and 288 on LBF (`target_rule`, `path_tiebreak`, `load_rule`). The noise levels were chosen off the values used by the held-out strangers so that no stranger is a grid member. A richer grammar with rule switching and seat swapping (up to 1,277 hypotheses on wide) changed nothing in the planning study and is not used.

Strangers. Table 1 lists the held-out strangers with their knobs. The four named defaults (`cook`, `flex`, `flex_conv1`, `signaller`) are noise-free, while the eight `held_*` partners and, on cramped room, the two `deliver_self` parkers carry noise rates and delays off the grid. The Overcooked list is the same on the three layouts except for the two parkers, which exist only on cramped room. The second, disjoint set (`train_sig_a/b`, `train_cook_a/b`, `train_flex_a/b`) was defined before the method existed, and no TTBR experiment used it until the final replication. On wide, partners with `reads_indicator` $= 1$ occupy the signaller seat and those with `reads_indicator` $= 0$ the cook seat. The ego takes the other seat, and on the symmetric layouts the ego alternates seats.

Blueprints, library and siblings. Blueprints are self-play PPO agents trained for 3×10^7 environment steps with the network and hyperparameters of Appendix D. Seeds 1000, 1001 and 1002 are the three blueprints of Table 2. The library $\mathcal{L}$ holds the checkpoints at 10%, 30% and 100% of training of eight self-play seeds (1000 to 1007), 24 entries, so the blueprint's own checkpoints are among them. The siblings are the final checkpoints of eight further self-play seeds (2000 to 2007), and the first four are the retention partners of every table and all eight are the neural strangers of RQ5.

Train-time baselines. SP is the blueprint scored as a method. FCP trains an ego against uniform samples from the 24-entry library, whose eight self-play members trained for 3×10^7 steps each with checkpoints at 10%, 30% and 100% of training. MEP trains eight population members in parallel, for the same 3×10^7 steps each, with the bonus $\alpha\, H(\bar{\pi})$ on the entropy of the population's average policy, exchanging parameters every ten updates, for $\alpha \in \{0.01, 0.05\}$, builds a 24-entry pool from their checkpoints at the same fractions, and then trains an FCP-style ego against it, so the ego stage uses uniform sampling rather than MEP's prioritised sampling, the simplification stated in §4.2. Every ego (FCP's, MEP's and the five variants') trains for 3×10^7 steps, the blueprint's own budget, with five seeds. The five population variants come from an earlier line of this project and differ in the ego's partner-input representation and auxiliary losses, and they are listed with the raw returns in Appendix E. The program-generalist uses the blueprint's architecture and budget and is trained against programs drawn uniformly from $\mathcal{P}$ on every seat, three seeds. TTBR from the generalist (RQ3) starts from the first seed. The generalist is recurrent and trained against a distribution of partners, so it also serves as the in-context baseline in the sense of RL^2 (Duan et al., 2016).

ICRL4AHT teammates. The external benchmark's 20 held-out heuristic teammates on the wide layout come from four stateful families (assembly line, territory, utility-greedy and recipe-aware) and run unmodified inside our evaluator through a host callback, and their randomness is a deterministic function of the decision history, as in the benchmark. Under the benchmark's protocol each teammate instance is scored over 100 episodes of 100 steps with greedy actions, and the published baseline numbers are copied from the benchmark's paper.

Table 1: Held-out strangers (first set) and the second set. Knobs not listed are at the family default.

name	knobs
`cook`	role cook, reads_indicator 0
`signaller`	role signaller, reads_indicator 1
`flex`, `flex_conv1`	role flexible, reads_indicator 1 (`flex_conv1`: signal_convention 1)
`held_sig_1`	reads_indicator 1, signal_convention 0, signal_delay 20, path_tiebreak 1, hesitation 0.2
`held_sig_2`	reads_indicator 1, signal_convention 1, path_tiebreak 0, wander 0.2
`held_sig_3`	reads_indicator 1, signal_convention 1, signal_delay 40, wrong action 0.1
`held_sig_4`	reads_indicator 1, signal_convention 0, hesitation 0.3, wander 0.3
`held_cook_1`	reads_indicator 0, plate_first 0, path_tiebreak 1, hesitation 0.2
`held_cook_2`	reads_indicator 0, signal_convention 1, plate_first 1, wander 0.2
`held_cook_3`	reads_indicator 0, signal_convention 1, plate_first 0, wrong action 0.1
`held_cook_4`	reads_indicator 0, plate_first 1, path_tiebreak 1, hesitation 0.3
`deliver_self:a/b`	deliver_self on / off pair (cramped room only)
`train_sig_a`	reads_indicator 1, signal_convention 0, path_tiebreak 0, hesitation 0.1
`train_sig_b`	reads_indicator 1, signal_convention 1, path_tiebreak 1, signal_delay 10, wander 0.1
`train_cook_a`	reads_indicator 0, plate_first 1, hesitation 0.1
`train_cook_b`	reads_indicator 0, signal_convention 1, plate_first 0, path_tiebreak 1, wander 0.1
`train_flex_a`	reads_indicator 1, hesitation 0.05, wrong action 0.05
`train_flex_b`	reads_indicator 1, signal_convention 1, path_tiebreak 1, wander 0.1, wrong action 0.05
LBF: `nearest`, `partner_nearest`, `farthest`, `lowest_id`	target_rule 0, 1, 2, 3
LBF: `held_*_1/2` (eight)	each target rule with path_tiebreak 0 / 1, load_rule and one noise knob at 0.1 or 0.2

C PROOFS

Throughout, $V(\pi, \tau)$ is finite for every pair, so every regret and conflict is a finite non-negative number, and $\mathrm{BR}(\tau_j)$ denotes the fixed best response chosen in Definition 1.

Proof of Proposition 1. By optimality of π_ω for the mixture, $\sum_k \omega_k V(\pi_\omega, \tau_k) \geq \sum_k \omega_k V(\mathrm{BR}(\tau_j), \tau_k)$. Separating the term $k = j$ and using $V(\mathrm{BR}(\tau_j), \tau_j) = V^\star_{\tau_j}$,

$$\begin{aligned} \omega_j \,\mathrm{Reg}_{\tau_j}(\pi_\omega) \;&\leq\; \sum_{k \neq j} \omega_k \big[V(\pi_\omega, \tau_k) - V(\mathrm{BR}(\tau_j), \tau_k) \big] \\ &\leq\; \sum_{k \neq j} \omega_k \big[V^\star_{\tau_k} - V(\mathrm{BR}(\tau_j), \tau_k) \big] \;=\; \sum_{k \neq j} \omega_k \, \kappa(j \to k), \end{aligned}$$

which is the first claim. For the second, $\max_\pi \sum_k \omega_k V(\pi, \tau_k) \geq \sum_k \omega_k V(\mathrm{BR}(\tau_j), \tau_k)$ for every j, so $\mathrm{VI}(\omega) \leq \sum_k \omega_k [V^\star_{\tau_k} - V(\mathrm{BR}(\tau_j), \tau_k)] = \sum_{k \neq j} \omega_k \, \kappa(j \to k)$, the $k = j$ term being zero, and minimising over j gives the bound. □

Proposition 2 (approximate maximisers)**.** *If π is an ε-approximate maximiser of the mixture objective, $\sum_k \omega_k V(\pi, \tau_k) \geq \max_{\pi'} \sum_k \omega_k V(\pi', \tau_k) - \varepsilon$, then for every j with $\omega_j > 0$, $\mathrm{Reg}_{\tau_j}(\pi) \leq \omega_j^{-1} \big[\sum_{k \neq j} \omega_k \, \kappa(j \to k) + \varepsilon \big]$.*

Proof. The maximum is at least the objective at $\mathrm{BR}(\tau_j)$, so $\sum_k \omega_k V(\pi, \tau_k) \geq \sum_k \omega_k V(\mathrm{BR}(\tau_j), \tau_k) - \varepsilon$, and the argument above goes through with ε carried along:

$\omega_j \operatorname{Reg}_{\tau_j}(\pi) \le \sum_{k\neq j} \omega_k [V(\pi,\tau_k) - V(\mathrm{BR}(\tau_j),\tau_k)] + \varepsilon \le \sum_{k\neq j} \omega_k \kappa(j \to k) + \varepsilon$. Applied to Theorem 1, an ε-approximate best response to the posterior has regret at most $\bar\kappa_\theta O_N + \varepsilon/b_N(\theta)$. □

Proof of Corollary 1. With $\kappa \equiv 0$ on the support, Proposition 1 gives $\operatorname{Reg}_{\tau_j}(\pi_\omega) \le 0$ for every j in the support, so π_ω is a best response to each such type, and $\mathrm{VI}(\omega) \le 0$, while $\mathrm{VI}(\omega) \ge 0$ holds by definition. Any ego that conditions on the interaction, however it does so, achieves at most $V^\star_{\tau_k}$ against each τ_k and hence at most $\sum_k \omega_k V^\star_{\tau_k}$ on average, which π_ω already attains. □

Proof of Lemma 1. Let h_t denote the full history of the interaction before the partner's t-th action, so that h_{t+1} contains h_t, the partner's action a_t, the ego's action and the next observations. Conditionally on h_t the partner draws $a_t \sim \tau_\theta(\cdot \mid h_t)$, where $\tau_\theta(\cdot \mid h_t)$ is the partner's action distribution given its own information, a function of h_t. Fix a rival $i \neq \theta$ and define

$$X_t = \sqrt{\tau_i(a_t \mid h_t)/\tau_\theta(a_t \mid h_t)}, \qquad \rho_t = \sum_a \sqrt{\tau_\theta(a \mid h_t)\,\tau_i(a \mid h_t)},$$

so that X_t is well defined almost surely and $\mathbb{E}[X_t \mid h_t] = \sum_a \tau_\theta(a \mid h_t)\sqrt{\tau_i(a \mid h_t)/\tau_\theta(a \mid h_t)} = \rho_t$, an h_t-measurable quantity in $[0,1]$ by the Cauchy–Schwarz inequality. Let $Y_t = X_t/\rho_t$ when $\rho_t > 0$ and $Y_t = 1$ when $\rho_t = 0$. In the latter case $\tau_i(a_t \mid h_t) = 0$ almost surely, hence $X_t = 0$. In both cases $\mathbb{E}[Y_t \mid h_t] = 1$ and $X_t = \rho_t Y_t$ almost surely.

From (1), $b_N(i)/b_N(\theta) = (p_i/p_\theta)\prod_{t\le N} \tau_i(a_t \mid h_t)/\tau_\theta(a_t \mid h_t)$, hence

$$\sqrt{b_N(i)/b_N(\theta)} \;=\; \sqrt{p_i/p_\theta}\,\prod_{t\le N} X_t \;=\; \sqrt{p_i/p_\theta}\; e^{-D_N(i)} \prod_{t\le N} Y_t,$$

with the convention $e^{-\infty} = 0$ when some $\rho_t = 0$, which is consistent because then $X_t = 0$. On the event $\{D_N(i) \ge d\}$ the factor $e^{-D_N(i)}$ is at most e^{-d}, and $\prod_t Y_t \ge 0$, so $\sqrt{b_N(i)/b_N(\theta)}\,\mathbf{1}\{D_N(i) \ge d\} \le \sqrt{p_i/p_\theta}\, e^{-d} \prod_{t\le N} Y_t$. Finally $\mathbb{E}[\prod_{t\le N} Y_t] = 1$: since $\prod_{t<N} Y_t$ is h_N-measurable, $\mathbb{E}[\prod_{t\le N} Y_t] = \mathbb{E}[\prod_{t<N} Y_t\, \mathbb{E}[Y_N \mid h_N]] = \mathbb{E}[\prod_{t<N} Y_t]$, and induction on N gives the claim. Taking expectations proves the first inequality. No property of the environment or of the ego's policy was used beyond their being part of h_t.

For the second inequality, $O_N = \sum_{i\neq\theta} b_N(i)/b_N(\theta)$ and $\sqrt{\sum_i x_i} \le \sum_i \sqrt{x_i}$ for non-negative x_i, so on $\mathcal{E}_d$, where every indicator $\mathbf{1}\{D_N(i) \ge d\}$ equals one, $\sqrt{O_N}\,\mathbf{1}\{\mathcal{E}_d\} \le \sum_{i\neq\theta} \sqrt{b_N(i)/b_N(\theta)}\,\mathbf{1}\{D_N(i) \ge d\}$. Take expectations and apply the first inequality to each term. □

Proof of Theorem 1. (i) Apply Proposition 1 with $\omega = b_N$ and $j = \theta$: $\operatorname{Reg}_{\tau_\theta}(\pi_b) \le b_N(\theta)^{-1} \sum_{k\neq\theta} b_N(k)\,\kappa(\theta \to k) \le \bar\kappa_\theta\,(1 - b_N(\theta))/b_N(\theta) = \bar\kappa_\theta\, O_N$. (ii) If $\bar\kappa_\theta = 0$ the event is empty by (i). Otherwise, by (i), the event $\{\mathcal{E}_d \text{ and } \operatorname{Reg}_{\tau_\theta}(\pi_b) > \bar\kappa_\theta S_\theta^2 e^{-2d}/\eta^2\}$ is contained in $\{\sqrt{O_N}\,\mathbf{1}\{\mathcal{E}_d\} > S_\theta e^{-d}/\eta\}$, whose probability is at most $\eta\,\mathbb{E}_\theta[\sqrt{O_N}\mathbf{1}\{\mathcal{E}_d\}]/(S_\theta e^{-d}) \le \eta$ by Markov's inequality and Lemma 1. □

Remark 1 (how much evidence suffices). With a uniform prior over n programs $S_\theta = n - 1$, so evidence of order $\log n$ nats suffices: a few dozen decisive observed steps for the classes of 72 to 288 programs used here.

Corollary 2 (the anchored objective)**.** *Let π_J maximise (2) and let θ be the partner's type. Write $\bar\kappa_\theta = \max_{k\neq\theta} \kappa(\theta \to k)$, $\bar\kappa_\theta(\mathcal{L}) = \max_{\ell\in\mathcal{L}} \kappa(\theta \to \ell)$, and $\kappa(\theta \to 0)$ for the conflict of θ towards the blueprint regarded as a partner. Then*

$$\operatorname{Reg}_{\tau_\theta}(\pi_J) \;\le\; \frac{\alpha_0\,\kappa(\theta \to 0) + \alpha_L\,\bar\kappa_\theta(\mathcal{L}) + \alpha_b\,(1 - b_N(\theta))\,\bar\kappa_\theta}{\alpha_b\, b_N(\theta)},$$

$$\operatorname{Reg}_{\pi_0}(\pi_J) \;\le\; \frac{\alpha_L\,\bar\kappa_0(\mathcal{L}) + \alpha_b \sum_i b_N(i)\,\kappa(0 \to i)}{\alpha_0}.$$

Proof. The objective (2) is a mixture over the component set $\{0\} \cup \mathcal{L} \cup \mathcal{H}$, where component 0 is the blueprint regarded as a partner, with weights $\omega_0 = \alpha_0$, $\omega_\ell = \alpha_L/|\mathcal{L}| + \alpha_b b_N(\ell)$ for $\ell \in \mathcal{L}$ (library

members are also hypotheses) and $\omega_i = \alpha_b b_N(i)$ for $i \in \mathcal{P}$. Apply Proposition 1 with $j = \theta$. The weight of θ is at least $\alpha_b b_N(\theta)$, and the numerator $\sum_{k\neq\theta} \omega_k \kappa(\theta \to k)$ is at most $\alpha_0\, \kappa(\theta \to 0) + \sum_\ell (\alpha_L/|\mathcal{L}|)\, \kappa(\theta \to \ell) + \alpha_b \sum_{k\neq\theta} b_N(k)\, \kappa(\theta \to k) \le \alpha_0\, \kappa(\theta \to 0) + \alpha_L\, \bar\kappa_\theta(\mathcal{L}) + \alpha_b(1 - b_N(\theta))\, \bar\kappa_\theta$, which gives the first bound. For the second, apply Proposition 1 with $j = 0$ and $\omega_0 = \alpha_0$. The numerator is $\sum_\ell (\alpha_L/|\mathcal{L}|)\, \kappa(0 \to \ell) + \alpha_b \sum_i b_N(i)\, \kappa(0 \to i) \le \alpha_L\, \bar\kappa_0(\mathcal{L}) + \alpha_b \sum_i b_N(i)\, \kappa(0 \to i)$. □

Remark 2 (the price of the anchors). Applied to the components of (2), the corollary says that the anchors cost stranger regret only through the conflict between the stranger's best response and in-class play, at an exchange rate set by the weights, so raising α_L buys in-class value and sells stranger gain. The floor sweep of Appendix F shows the trade empirically. Two consequences are worth stating. First, the anchors set a floor that identification cannot remove: even at $b_N(\theta) = 1$ the stranger bound reads $[\alpha_0\, \kappa(\theta \to 0) + \alpha_L\, \bar\kappa_\theta(\mathcal{L})]/\alpha_b$, that is $(\kappa(\theta \to 0) + \bar\kappa_\theta(\mathcal{L}))/2$ for the recipe's weights, which is the price of retention. Second, the deployed policy is not the exact maximiser of (2) but a PPO fine-tune, and by Proposition 2 an ε-approximate maximiser satisfies the first bound with ε added to its numerator, an extra $\varepsilon/(\alpha_b\, b_N(\theta))$. Nothing in this paper bounds the optimisation error ε; the budget curve of Appendix F is its empirical trace.

Remark 3 (the averaged posterior). The adaptation phase samples from $\bar b = E^{-1} \sum_e b^{(e)}$, the average of the per-episode posteriors of the observation phase (Algorithm 1). Its odds against the truth satisfy $\bar O = \sum_e (1 - b_\theta^{(e)}) / \sum_e b_\theta^{(e)} \le \max_e O^{(e)}$ by the mediant inequality, so on the event that every episode accumulated evidence d against every rival, $\mathbb{E}[\sqrt{\bar O}\, \mathbf{1}\{\cdot\}] \le \sum_e \mathbb{E}[\sqrt{O^{(e)}}\, \mathbf{1}\{\mathcal{E}_d^{(e)}\}] \le E\, S_\theta e^{-d}$: Lemma 1 and Theorem 1 hold for the averaged posterior with S_θ replaced by $E S_\theta$. The averaging trades a factor E in the bound for robustness to a single uninformative episode.

Remark 4 (on the in-class bound). If the blueprint is a best response to itself, $\kappa(0 \to i) = \mathrm{Reg}_{\tau_i}(\pi_0)$ is the blueprint's own regret against type i, which is large for the strangers of this paper. The in-class bound of Corollary 2 is then loose and says only that in-class loss is at most $(1 - \alpha_0)/\alpha_0$ times the average stranger regret. What we observe is an in-class *gain* with the library floor (Appendix F), which the bound permits but does not predict.

Remark 5 (misspecification). If the partner is outside $\mathcal{H}$, the odds between two hypotheses i, j still evolve by the likelihood ratio $\prod_t \tau_i(a_t \mid h_t)/\tau_j(a_t \mid h_t)$ under the true partner. In expectation the log-odds drift towards the hypothesis with the smaller accumulated Kullback–Leibler divergence from the true behaviour along the visited histories (Berk, 1966). The response to that hypothesis has a regret that depends on the true partner's best response, which nothing in the observation reveals. Theorem 1 therefore says nothing about partners outside the class, and §4 (RQ5) reports them separately. The strangers of §4.1 are in-family but off-grid, so their nearest hypothesis is a program with neighbouring knob values. The next corollary says what that costs.

Corollary 3 (an imperfect class)**.** *Let the partner be any policy $\tau^\star$, and suppose some hypothesis $\tau_\theta \in \mathcal{H}$ is an ε-proxy of it: $|V(\pi, \tau^\star) - V(\pi, \tau_\theta)| \le \varepsilon$ for every ego policy π. Then* $\mathrm{Reg}_{\tau^\star}(\pi) \le \mathrm{Reg}_{\tau_\theta}(\pi) + 2\varepsilon$ *for every π, and for the best response π_b to any belief b_N with $b_N(\theta) > 0$,*

$$\mathrm{Reg}_{\tau^\star}(\pi_b) \;\le\; b_N(\theta)^{-1} \sum_{k\neq\theta} b_N(k)\, \kappa(\theta \to k) + 2\varepsilon \;\le\; \bar\kappa_\theta\, O_N(\theta) + 2\varepsilon,$$

where $O_N(\theta) = (1 - b_N(\theta))/b_N(\theta)$ are the odds the observation left against the proxy.

Proof. $V^\star_{\tau^\star} = \max_\pi V(\pi, \tau^\star) \le \max_\pi V(\pi, \tau_\theta) + \varepsilon = V^\star_{\tau_\theta} + \varepsilon$ and $V(\pi, \tau^\star) \ge V(\pi, \tau_\theta) - \varepsilon$, so $\mathrm{Reg}_{\tau^\star}(\pi) \le \mathrm{Reg}_{\tau_\theta}(\pi) + 2\varepsilon$. Part (i) of Theorem 1 is Proposition 1 applied to the belief as a set of weights, and it holds for every index of positive weight whether or not the belief was formed under that index, which gives the second claim. □

Remark 6 (reading the corollary). The corollary separates the two ways an imperfect class can hurt. The model error enters additively, and the oracle arm of RQ3 measures it for our off-grid strangers: the true-program fine-tune improves on the posterior response by at most $+6.6$. The odds against the proxy are a quantity the agent can read off its own posterior at test time, but nothing guarantees them small: Lemma 1 assumed that τ_θ generated the observation, and under $\tau^\star$ the posterior drifts towards the hypothesis nearest to the true behaviour in Kullback–Leibler divergence along the visited histories, which need not be the proxy nearest in value. This is the regime of the ICRL4AHT teammates (RQ5), where the response to the nearest hypothesis does no harm but has no edge.

Remark 7 (the likelihood floor and the planner's prior). The implementation replaces every likelihood $\tau_i(a_t \mid h_t)$ in (1) by $\max\{\tau_i(a_t \mid h_t), \epsilon\}$ with $\epsilon = 10^{-9}$, so that no hypothesis is driven to exactly zero by a single step. The proof of Lemma 1 goes through verbatim for the floored rival with ρ_t replaced by $\rho_t^\epsilon = \sum_a \sqrt{\tau_\theta(a \mid h_t)\, \max\{\tau_i(a \mid h_t), \epsilon\}} \le \rho_t + \sqrt{\epsilon|\mathcal{A}|}$, since only $\mathbb{E}[X_t \mid h_t] = \rho_t^\epsilon$ was used (the inequality is $\sqrt{x+\epsilon} \le \sqrt{x} + \sqrt{\epsilon}$ followed by Cauchy–Schwarz), and flooring the true type's own likelihood only lowers the odds against it, so Lemma 1 and Theorem 1 hold for the implemented filter with D_N computed from ρ_t^ϵ. The floor keeps the evidence of a single step finite, at most $\frac{1}{2}\log(1/\epsilon) \approx 10.4$ nats, and a step at which the two predictions are disjoint still contributes at least $\frac{1}{2}\log(\epsilon|\mathcal{A}|)^{-1} \approx 9.5$ nats for $|\mathcal{A}| = 6$, more than the $\log S_\theta < 6$ nats of Remark 1, while no hypothesis is eliminated by a single wrong action, which is the floor's purpose. The planner mixes the averaged belief $\bar{b}$ with 1% of the default prior (Appendix D), which adds at most $1/(99\,\bar{b}(\theta))$ to the odds against the truth.

D Implementation and compute

Algorithm 1 Test-time best response (TTBR)

Require: blueprint π_0; hypothesis class $\mathcal{H} = \mathcal{P} \cup \mathcal{L}$ with prior p; weights $(\alpha_0, \alpha_L, \alpha_b)$; budget U; simulator

1: **Observe:** for $e = 1, \dots, E$ (on every seat the partner can take): play one episode of π_0 with the partner, record the partner's actions and the states, and filter $b^{(e)}$ by (1)
2: $b \leftarrow \frac{1}{E}\sum_e b^{(e)}$
3: **Adapt:** $\pi \leftarrow \pi_0$
4: **for** $u = 1, \dots, U$ **do**
5: in each parallel environment draw a kind from $(\alpha_0, \alpha_L, \alpha_b)$ and a partner: π_0, a uniform $\ell \in \mathcal{L}$, or $i \sim b$
6: collect one rollout per environment and apply a PPO update to π
7: **end for**
8: **Check** (post-hoc addition, Appendix F): if the return against the belief never rose during adaptation, $\pi \leftarrow \pi_0$
9: **Deploy** π (optionally with a planner on top, §4.2)

Network and PPO. Every learned agent uses the same actor–critic: a grid encoder (1×1 convolutions of 128, 128 and 8 channels, then 3×3 convolutions of 16, 32 and 32 channels, a dense layer to 128 units and layer normalisation), a GRU with 128 hidden units, and linear policy and value heads. PPO (Schulman et al., 2017) runs 256 parallel environments for 256 steps per update (65,536 steps), four epochs of 64 minibatches, clipping 0.2 for the policy and the value, $\gamma = 0.99$, $\lambda = 0.95$, entropy coefficient 0.01, value coefficient 0.5, gradient norm clipped at 0.25 and a learning rate of 2.5×10^{-4} with linear warm-up over the first 5% of updates and linear annealing to zero, and reward shaping is annealed over the first half of training. Blueprints train for 3×10^7 steps (≈ 458 updates).

Observation and belief. The observation phase runs the blueprint with the partner for $E = 16$ episodes per seat (32 on LBF) with the planning evaluator's search disabled, and the union model filters (1) along each episode with prior mass $\rho = 0.5$ on the programs, likelihood temperature 1 and a likelihood floor of 10^{-9} (so that a hypothesis is not driven to exactly zero by a single step), and the final per-episode posteriors are averaged over episodes and seats. From the averaged belief the mixture of (2) is formed as: self-play $\alpha_0 = 0.25$; library $\alpha_L + 0.5\, m_L$ where m_L is the posterior mass on the library, with library members weighted by $0.5\, m_L\, b(\ell) + \alpha_L/|\mathcal{L}|$; programs $0.5\,(1 - m_L)$ with weights proportional to the posterior restricted to the programs.

Adaptation. Fine-tuning restarts the PPO recipe above from the blueprint's parameters with a fresh optimiser state and a fresh learning-rate schedule, for $U = 100$ updates (6,553,600 steps). Each parallel environment draws its partner kind and partner at reset. A fine-tune takes about 70 minutes on eight CPU cores of the cluster (43 seconds per update), and $U = 50$ takes about 38 minutes. The deployment check compares the mean of the last quarter of per-update returns against the belief with the mean of the first quarter.

Evaluation. Policies are scored by the trainer-side evaluator with 32 episodes per partner per seat and evaluation seed 0 for every policy, so that the blueprint's row and the adapted row see the same episode keys and every gain is paired. Planning arms are scored by the planning evaluator on 16 paired environments with the same partner list. The planner uses one-ply lookahead with N rollouts per candidate action ($N \in \{8, 16, 32, 64\}$, 8 unless stated), horizon $H \in \{8, 16, 32, 64\}$, discount 0.99, no margin, and the blueprint's policy for the ego inside rollouts, while partners inside rollouts are drawn from the partner model. Planning with TTBR's posterior loads the averaged belief as the prior, mixed with 1% of the default prior so that no hypothesis starts at exactly zero. Planning arms are scored on 16 paired environments, and means over strangers are taken over the two fine-tuning seeds first where they exist. Two conventions hold throughout the paper: on cramped room the main table uses all 14 strangers (mean gain $+69.7$) while comparisons with planning use the 12 common to all Overcooked layouts ($+59.7$ for the same policies), and since some strangers behave identically on a layout (of the 12 common strangers, 9 are distinct on cramped room and asymmetric advantages and 11 on wide; the 14 of cramped room hold 10) intervals over strangers slightly overstate the independent units. Three ablated planners bracket the posterior planner of RQ4: planning with the stranger's exact program (the oracle planner), planning with the library alone as the hypothesis class, and planning with a posterior filtered only within the scored episode (no observation phase).

Compute. Every run used CPU nodes of a Slurm cluster. One TTBR adaptation is 6.6×10^6 simulator steps, and the whole study (blueprints, populations, baselines, 58 adaptation items per blueprint with two fine-tuning seeds, planning arms at four horizons and the cross tables) is of the order of 10^4 core-hours. Planning at $H = 32$ with eight rollouts costs about 0.6×10^6 simulator steps per 400-step episode and 1.2×10^6 at $H = 64$.

E FULL RESULT TABLES

Table 2 summarises the main result per layout and blueprint, Table 3 the train-time baselines and the conflict fractions, and Table 4 the neural strangers and the external benchmark. The tables after them give the per-stranger values behind every mean in the main text: the main result per stranger and per blueprint, the identification ladder per stranger, the second stranger set, the planning tiers, the neural strangers per seed, the ICRL4AHT instances, and the train-time baselines with the five population variants named. Figure 7 shows the four cross tables of RQ1. Two conventions of the captions: every return is measured by the trainer-side evaluator of Appendix D (32 episodes per partner per seat, evaluation seed 0), called `fed_reeval` in the released code, and the identifiers E1, E2, E3, E4, G2, G4, G8 and W7-D name the experiments of the pre-registration record (Tables 28 and 29) whose stored results the table reads. Strangers that behave identically on a layout produce identical rows (`cook`, `flex`, `flex_conv1` and `cramped_deliver_self:a` on cramped room, `held_sig_2` and `held_cook_2` on cramped room and asymmetric advantages, `flex` and `signaller` on wide, and the five utility-greedy instances of ICRL4AHT), and they are counted once in the intervals over distinct behaviours of Table 2.

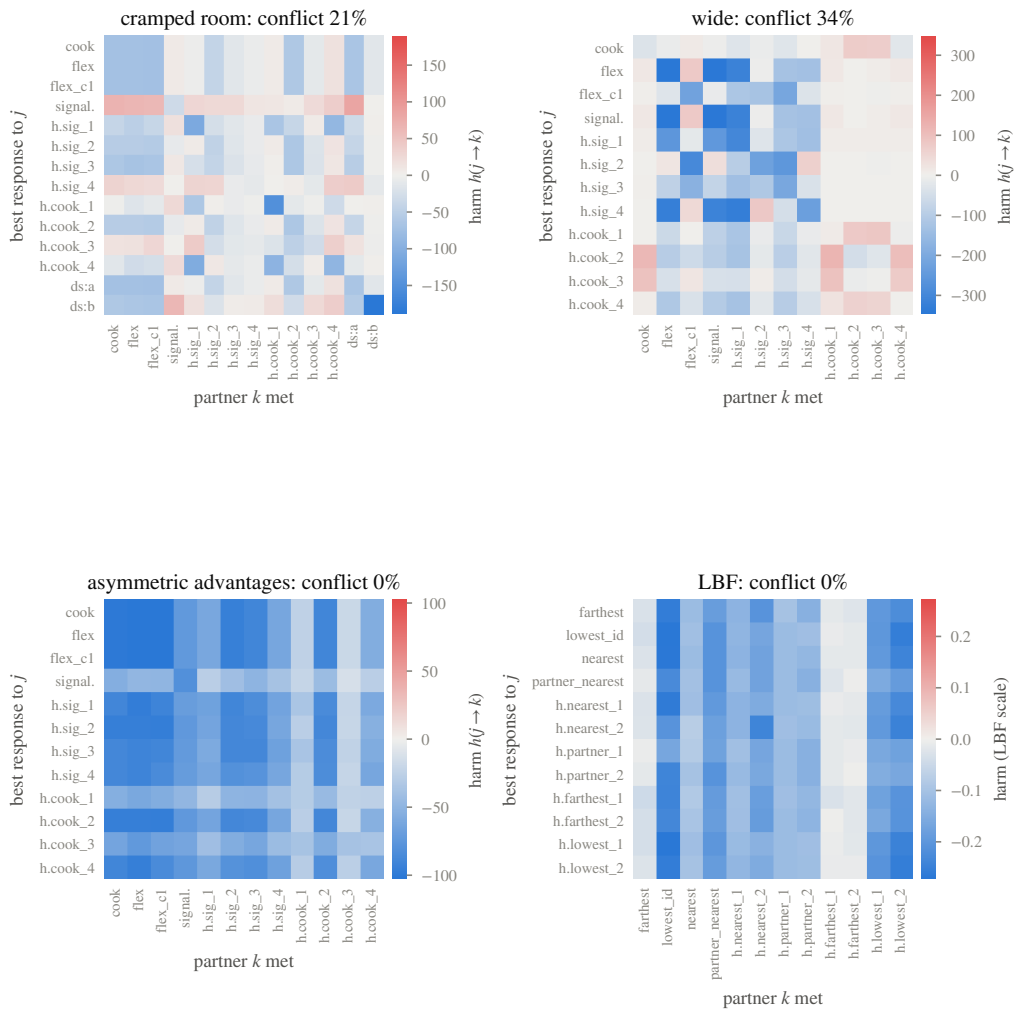

Figure 7: **Cross tables of best responses.** Harm $h(j \to k) = V(\pi_0, \tau_k) - V(\mathrm{BR}(\tau_j), \tau_k)$ when the best response fine-tuned for stranger j (row) meets stranger k (column). Red is worse than not adapting, blue is better. Gray cells on wide are pairs on different seats, which are not counted. The titles give the fraction of same-seat pairs with $h > 10$.

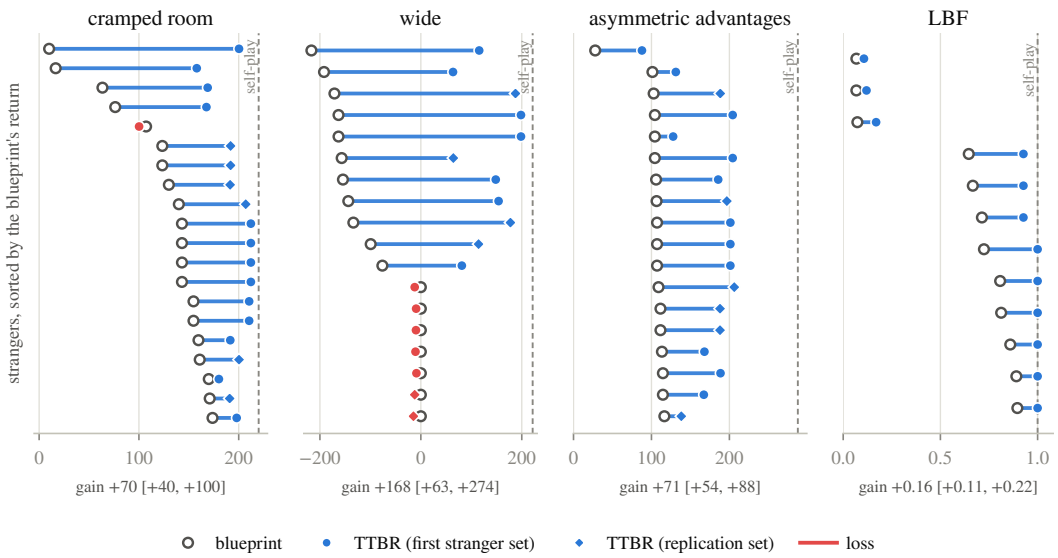

Figure 8: **Main result per stranger** (blueprint 1000). Hollow: blueprint, filled: adapted policy, diamonds: replication set, dashed: self-play return. Below each panel, the mean gain over the first set with its 95% interval.

Table 2: **Main result.** Mean return of the blueprint and of TTBR against the held-out strangers of each layout, the paired gain with its 95% interval over strangers and, beside it, over distinct behaviours (strangers that behave identically on the layout merged into one unit: 10 units among the 14 of cramped room, 11 of 12 on wide, 9 of 12 on asymmetric advantages, none merged on LBF), the number of strangers with a positive gain, and retention as the mean change against four held-out siblings with, in brackets, the worst fine-tune's mean change over the same four siblings (per-sibling values are not stored); the registered floor of -10 binds every fine-tune on cramped room and the mean over fine-tunes on wide. Three self-play blueprints per layout. The second block is the disjoint six-stranger set (registered verdicts PASS, INCONCLUSIVE, PASS) and the last block pools both sets. The asymmetric-advantages returns of blueprints 1001 and 1002 are rounded from the run logs, their intervals were computed from the stored per-stranger gains after the fact and were not registered, and blueprint 1002 there reads NOT REPLICATED under the registered rule ($+29.4 < 0.5 \times 71.3$). Every number is ungated: the deployment check of Algorithm 1 would move the wide mean of blueprint 1000 to $+172.4$ and change nothing else, so the reported values are the conservative ones.

layout (strangers, self-play return)	blueprint	blueprint	TTBR	gain [95% CI], strangers	distinct behaviours	positive	retention (worst)
cramped room (14, 220)	1000	118.5	188.2	+69.7 [+39.8, +99.5]	+71.3 [+27.0, +115.6]	13/14	+24.8 (+16.1)
	1001	64.7	186.9	+122.2 [+82.0, +162.4]	+99.0 [+48.9, +149.1]	14/14	+6.5 (−0.9)
	1002	69.6	175.9	+106.3 [+72.9, +139.7]	+92.5 [+48.1, +136.8]	13/14	+0.7 (−5.1)
wide (12, 222)	1000	−92.4	75.8	+168.2 [+62.7, +273.8]	+150.7 [+41.8, +259.5]	7/12	+31.3 (+16.9)
	1001	−85.6	89.3	+174.9 [+66.8, +283.0]	+157.9 [+45.4, +270.4]	8/12	+24.7 (+9.0)
	1002	−40.7	92.4	+133.1 [+48.3, +217.9]	+118.0 [+31.6, +204.4]	7/12	−6.1 (**−17.1**)
asymmetric adv. (12, 288)	1000	101.0	172.4	+71.3 [+54.2, +88.5]	+63.1 [+42.8, +83.4]	12/12	+26.6 (+19.0)
	1001	140	181	+40.5 [+31.9, +49.1]	+36.9 [+26.2, +47.6]	12/12	+2.0 (−1.1)
	1002	151	181	+29.4 [+21.4, +37.4]	+26.9 [+16.4, +37.5]	12/12	−3.9 (−7.3)
LBF (12, 1.0)	1000	0.62	0.77	+0.163 [+0.109, +0.217]	same	12/12	0.00 (0.00)
	1001	0.61	0.77	+0.156 [+0.101, +0.211]	same	12/12	0.00 (0.00)
	1002	0.58	0.76	+0.183 [+0.122, +0.244]	same	12/12	0.00 (0.00)
second set: cramped room (6)	1000	141.5	195.6	+54.1 [+33.1, +75.2]	+51.2 [+25.2, +77.3]	6/6	–
second set: wide (6)	1000	−93.5	86.1	+179.7 [+12.5, +346.9]	same	4/6	–
second set: asymmetric adv. (6)	1000	109.7	184.2	+74.6 [+46.1, +103.0]	+74.2 [+36.6, +111.8]	6/6	–
both sets: cramped room (20)	1000	–	–	+65.0 [+44.1, +85.9]	–	19/20	–
both sets: wide (18)	1000	–	–	+172.0 [+92.9, +251.2]	–	11/18	–
both sets: asymmetric adv. (18)	1000	–	–	+72.4 [+59.3, +85.5]	–	18/18	–

Table 3: **Train-time diversity against the strangers, and the conflict behind it.** Left: mean return over the full stranger set of each layout, 14 on cramped room and 12 on wide, every policy scored with the same instrument (32 episodes per partner per seat, the blueprint's self-play return at 220 on cramped room and 222 on wide, five ego seeds per trained method, and MEP with the better of two bonus weights). Right: from 400 cross evaluations of every stranger's best response against every other stranger of its layout, the fraction of same-seat pairs with harm $h > 10$, the share of the per-partner fine-tunes' gains that the best single fine-tune recovers when it serves every stranger, and the worst pair. On asymmetric advantages and LBF the fraction is 0% and the share 0.97 and 0.93.

layout	SP (blueprint)	FCP	variants	MEP	generalist	TTBR	conflict	one resp.	worst
cramped room	121	75	71–76	58	152	**188**	21%	0.48	−83
wide	−78	−18	−12 . . . − 8	−8	27	**76**	34%	0.59	−121

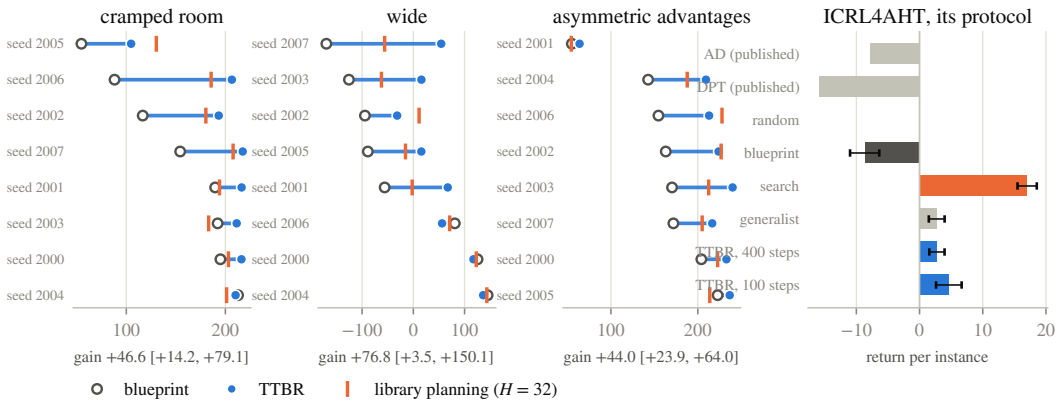

Figure 9: **Beyond scripted partners.** Left: return with each held-out self-play seed, blueprint (hollow) to TTBR (filled), with planning with the library at $H = 32$ as a tick and the mean gain below. Right: the ICRL4AHT teammates under the benchmark's protocol.

Table 4: **Beyond scripted partners.** Left: mean gain against eight held-out self-play seeds (95% interval over seeds, positive count, two fine-tuning seeds) for TTBR, for planning with the library at the default and the long horizon, and for a planner handed the partner's own network at $H = 32$. Right: the ICRL4AHT heuristic teammates on wide under the benchmark's protocol (100 episodes of 100 steps, greedy, published baselines from the benchmark) and under ours.

layout	TTBR	library $H=8$	library $H=32$	own network
cramped room	+46.6 [+14.2, +79.1] (7/8)	+46.0	+35.3	+52.1
wide	+76.8 [+3.5, +150.1] (5/8)	+42.2	+49.5	+100.5
asymmetric adv.	+44.0 [+23.9, +64.0] (8/8)	+16.2	+33.3	+50.9

ICRL4AHT, their protocol	return
AD / DPT / random	−7.9 / −15.9 / 0.0
blueprint	−8.7
search	+17.0
TTBR (adapted at 400 / 100 steps)	+2.7 / +4.6
generalist	+2.7
our protocol: TTBR gain	+173.7 [+148.0, +199.3]

Stranger	Blueprint	TTBR	Gain	Retention
cook	143.1	212.2	+69.1	+27.9
flex	143.1	212.2	+69.1	+27.9
flex_conv1	143.1	212.2	+69.1	+27.9
signaller	107.2	100.2	−7.0	+22.8
held_sig_1	63.4	168.9	+105.5	+16.1
held_sig_2	154.7	210.5	+55.8	+25.6
held_sig_3	173.8	198.0	+24.2	+30.5
held_sig_4	170.0	180.2	+10.2	+27.9
held_cook_1	16.6	158.0	+141.4	+20.7
held_cook_2	154.7	210.5	+55.8	+25.6
held_cook_3	159.7	191.6	+31.9	+21.4
held_cook_4	76.2	167.7	+91.4	+19.6
cramped_deliver_self:a	143.1	212.2	+69.1	+27.9
cramped_deliver_self:b	10.0	200.3	+190.3	+25.5

Table 5: Per-stranger result on *Cramped Room* for blueprint 1000 (final recipe), $n = 14$ strangers, of which four (cook, flex, flex_conv1, cramped_deliver_self:a) and two more (held_sig_2, held_cook_2) behave identically on this layout and give identical rows. Blueprint and TTBR columns are fed_reeval returns (32 episodes/partner/seat), each a point estimate averaged over the (up to 2) available fine-tune seeds for that stranger, Gain is TTBR − Blueprint and Retention is the mean change on the 4 held-out self-play sibling seeds under the same fine-tune. No confidence interval is stored at the per-stranger level, and the pooled 95% t-interval over the 14 strangers is reported in the main text.

Stranger	Blueprint	TTBR	Gain	Retention
cook	0.0	−12.2	−12.2	+41.2
flex	−163.1	198.4	+361.6	+26.2
flex_conv1	−154.4	148.4	+302.8	+20.2
signaller	−163.1	198.4	+361.6	+26.2
held_sig_1	−216.9	115.6	+332.5	+27.9
held_sig_2	−76.2	81.2	+157.5	+16.9
held_sig_3	−191.9	63.8	+255.6	+27.5
held_sig_4	−143.8	154.1	+297.8	+38.9
held_cook_1	0.0	−9.4	−9.4	+39.3
held_cook_2	0.0	−9.7	−9.7	+41.2
held_cook_3	0.0	−10.6	−10.6	+38.2
held_cook_4	0.0	−8.8	−8.8	+32.5

Table 6: Per-stranger result on *Wide* for blueprint 1000 (final recipe), $n = 12$ strangers (flex and signaller behave identically on this layout). Blueprint and TTBR columns are fed_reeval returns (32 episodes/partner/seat), each a point estimate averaged over the (up to 2) available fine-tune seeds for that stranger, Gain is TTBR − Blueprint and Retention is the mean change on the 4 held-out self-play sibling seeds under the same fine-tune. No confidence interval is stored at the per-stranger level, and the pooled 95% t-interval over the 12 strangers is reported in the main text.

Stranger	Blueprint	TTBR	Gain	Retention
cook	107.2	201.2	+94.1	+21.2
flex	107.2	201.2	+94.1	+21.2
flex_conv1	107.2	201.2	+94.1	+21.2
signaller	27.8	87.8	+60.0	+34.6
held_sig_1	105.9	185.6	+79.7	+31.2
held_sig_2	104.4	204.2	+99.8	+21.8
held_sig_3	114.7	188.6	+73.9	+35.5
held_sig_4	113.4	168.0	+54.5	+37.0
held_cook_1	104.4	127.8	+23.4	+19.0
held_cook_2	104.4	204.2	+99.8	+21.8
held_cook_3	101.2	131.2	+30.0	+21.3
held_cook_4	114.7	167.2	+52.5	+34.0

Table 7: Per-stranger result on *Asymm* for blueprint 1000 (final recipe), $n = 12$ strangers (cook, flex, flex_conv1 and, separately, held_sig_2, held_cook_2 behave identically on this layout). Blueprint and TTBR columns are fed_reeval returns (32 episodes/partner/seat), each a point estimate averaged over the (up to 2) available fine-tune seeds for that stranger, Gain is TTBR − Blueprint and Retention is the mean change on the 4 held-out self-play sibling seeds under the same fine-tune. No confidence interval is stored at the per-stranger level, and the pooled 95% t-interval over the 12 strangers is reported in the main text.

Stranger	Blueprint	TTBR	Gain	Retention
farthest	0.068	0.107	+0.039	0.000
lowest_id	0.646	0.927	+0.281	0.000
nearest	0.807	1.000	+0.193	0.000
partner_nearest	0.812	1.000	+0.188	0.000
held_nearest_1	0.859	1.000	+0.141	0.000
held_nearest_2	0.724	1.000	+0.276	0.000
held_partner_1	0.891	1.000	+0.109	0.000
held_partner_2	0.896	1.000	+0.104	0.000
held_farthest_1	0.068	0.120	+0.052	−0.001
held_farthest_2	0.073	0.169	+0.096	0.000
held_lowest_1	0.714	0.927	+0.214	0.000
held_lowest_2	0.667	0.927	+0.260	0.000

Table 8: Per-stranger result on *LBF* for blueprint 1000 (final recipe), $n = 12$ strangers. Blueprint and TTBR columns are fed_reeval returns (32 episodes/partner/seat), each a point estimate averaged over the (up to 2) available fine-tune seeds for that stranger, Gain is TTBR − Blueprint and Retention is the mean change on the 4 held-out self-play sibling seeds under the same fine-tune. No confidence interval is stored at the per-stranger level, and the pooled 95% t-interval over the 12 strangers is reported in the main text.

Stranger	bp1000 gain	bp1001 gain	bp1002 gain
cook	+69.1	+180.9	+154.7
flex	+69.1	+180.9	+154.7
flex_conv1	+69.1	+180.9	+154.7
signaller	−7.0	+11.6	−0.9
held_sig_1	+105.5	+100.3	+115.0
held_sig_2	+55.8	+177.8	+99.4
held_sig_3	+24.2	+61.9	+42.8
held_sig_4	+10.2	+10.0	+18.8
held_cook_1	+141.4	+137.5	+125.6
held_cook_2	+55.8	+177.8	+99.4
held_cook_3	+31.9	+24.1	+49.1
held_cook_4	+91.4	+101.9	+133.8
cramped_deliver_self:a	+69.1	+180.9	+154.7
cramped_deliver_self:b	+190.3	+184.1	+186.6

Table 9: Per-stranger TTBR gain (TTBR − blueprint, paired by stranger) on *Cramped Room* for three independently trained self-play blueprints, $n = 14$ strangers. Each cell is a single point estimate (fed_reeval, 32 episodes/partner/seat), and pooled 95% t-intervals per blueprint are reported in the main text.

Stranger	bp1000 gain	bp1001 gain	bp1002 gain
cook	−12.2	−8.8	−3.1
flex	+361.6	+361.9	+300.0
flex_conv1	+302.8	+360.0	+197.5
signaller	+361.6	+361.9	+300.0
held_sig_1	+332.5	+335.6	+295.6
held_sig_2	+157.5	+204.4	+216.2
held_sig_3	+255.6	+268.8	+154.4
held_sig_4	+297.8	+258.1	+175.6
held_cook_1	−9.4	−21.2	−5.6
held_cook_2	−9.7	−16.2	−2.5
held_cook_3	−10.6	−8.1	−23.8
held_cook_4	−8.8	+2.5	−6.9

Table 10: Per-stranger TTBR gain (TTBR − blueprint, paired by stranger) on *Wide* for three independently trained self-play blueprints, $n = 12$ strangers. Each cell is a single point estimate (fed_reeval, 32 episodes/partner/seat), and pooled 95% t-intervals per blueprint are reported in the main text.

Stranger	bp1000 gain	bp1001 gain	bp1002 gain
cook	+94.1	+50.3	+35.0
flex	+94.1	+50.3	+35.0
flex_conv1	+94.1	+50.3	+35.0
signaller	+60.0	+10.3	+4.1
held_sig_1	+79.7	+48.4	+34.7
held_sig_2	+99.8	+52.8	+40.6
held_sig_3	+73.9	+47.5	+45.6
held_sig_4	+54.5	+29.1	+18.1
held_cook_1	+23.4	+30.3	+23.4
held_cook_2	+99.8	+52.8	+40.6
held_cook_3	+30.0	+34.1	+11.9
held_cook_4	+52.5	+29.7	+29.1

Table 11: Per-stranger TTBR gain (TTBR − blueprint, paired by stranger) on *Asymm* for three independently trained self-play blueprints, $n = 12$ strangers. Each cell is a single point estimate (fed_reeval, 32 episodes/partner/seat), and pooled 95% t-intervals per blueprint are reported in the main text.

Stranger	bp1000 gain	bp1001 gain	bp1002 gain
farthest	+0.039	+0.042	+0.068
lowest_id	+0.281	+0.250	+0.344
nearest	+0.193	+0.188	+0.214
partner_nearest	+0.188	+0.193	+0.276
held_nearest_1	+0.141	+0.167	+0.151
held_nearest_2	+0.276	+0.276	+0.255
held_partner_1	+0.109	+0.094	+0.115
held_partner_2	+0.104	+0.073	+0.156
held_farthest_1	+0.052	+0.047	+0.063
held_farthest_2	+0.096	+0.073	+0.063
held_lowest_1	+0.214	+0.214	+0.286
held_lowest_2	+0.260	+0.255	+0.208

Table 12: Per-stranger TTBR gain (TTBR − blueprint, paired by stranger) on *LBF* for three independently trained self-play blueprints, $n = 12$ strangers. Each cell is a single point estimate (fed_reeval, 32 episodes/partner/seat), and pooled 95% t-intervals per blueprint are reported in the main text.

Stranger	Blueprint	Prior BR	Generalist	MAP	Posterior	True prog.
cook	143.1	79.8	178.4	213.1	212.2	215.5
flex	140.0	93.6	185.8	213.1	212.2	215.5
flex_conv1	138.1	98.1	180.2	213.1	212.2	215.5
signaller	112.2	56.4	92.5	112.5	100.2	143.4
held_sig_1	72.2	49.1	63.6	175.8	168.9	181.1
held_sig_2	167.2	182.7	202.0	212.5	210.5	213.8
held_sig_3	178.8	178.3	193.9	198.4	198.0	193.4
held_sig_4	172.2	171.9	182.9	183.1	180.2	181.9
held_cook_1	25.3	21.6	86.8	177.5	158.0	170.8
held_cook_2	147.8	184.7	202.0	212.5	210.5	213.8
held_cook_3	160.6	172.7	188.8	193.0	191.6	190.0
held_cook_4	78.4	45.5	55.3	170.8	167.7	173.1
cramped_deliver_self:a	148.4	90.3	181.5	213.1	212.2	215.5
cramped_deliver_self:b	10.0	40.0	139.4	195.2	200.3	204.1

Table 13: Identification ladder (E1) per stranger on *Cramped Room*, $n = 14$ strangers: mean return (fed_reeval) of the blueprint, a best response to the uniform prior at $U = 100$, the converged program-generalist, the best response to the posterior's MAP program, the full posterior of the observation phase (= TTBR), and a best response to the partner's true program. Single point estimate per stranger from the E1 evaluation tree, and this tree's own blueprint column differs by a few points from Table 5's (two different fed_reeval evaluation trees for the same checkpoint). Paired-gain 95% t-intervals (R1–R4) over these strangers are reported in the main text.

Stranger	Blueprint	Prior BR	Generalist	MAP	Posterior	True prog.
cook	0.0	−19.4	−21.7	−2.8	−12.2	33.4
flex	−155.6	24.1	60.8	210.3	198.4	195.3
flex_conv1	−118.8	22.8	54.6	70.3	148.4	91.9
signaller	−145.0	28.8	62.9	210.3	198.4	195.3
held_sig_1	−198.8	30.9	59.8	82.2	115.6	71.2
held_sig_2	−63.1	18.1	57.7	83.4	81.2	166.2
held_sig_3	−159.4	20.0	57.1	105.3	63.8	33.4
held_sig_4	−88.1	28.8	44.0	151.9	154.1	132.2
held_cook_1	0.0	−35.6	−12.1	−24.4	−9.4	−12.5
held_cook_2	0.0	−13.8	−16.2	−71.6	−9.7	60.0
held_cook_3	−3.1	−19.7	−12.3	−58.4	−10.6	3.8
held_cook_4	0.0	−20.3	−5.8	−1.2	−8.8	3.1

Table 14: Identification ladder (E1) per stranger on *Wide*, $n = 12$ strangers: mean return (fed_reeval) of the blueprint, a best response to the uniform prior at $U = 100$, the converged program-generalist, the best response to the posterior's MAP program, the full posterior of the observation phase (= TTBR), and a best response to the partner's true program. Single point estimate per stranger from the E1 evaluation tree, and this tree's own blueprint column differs by a few points from Table 6's (two different `fed_reeval` evaluation trees for the same checkpoint). Paired-gain 95% t-intervals (R1–R4) over these strangers are reported in the main text.

Stranger	Blueprint	Prior BR	Generalist	MAP	Posterior	True prog.
cook	107.2	193.6	205.2	200.2	201.2	207.5
flex	102.5	185.8	206.1	200.2	201.2	207.5
flex_conv1	105.9	186.1	203.5	200.2	201.2	207.5
signaller	27.2	100.2	96.4	94.4	87.8	106.1
held_sig_1	117.2	175.0	188.1	182.0	185.6	179.7
held_sig_2	111.6	190.2	206.1	201.7	204.2	204.2
held_sig_3	100.3	180.9	196.0	189.7	188.6	191.6
held_sig_4	110.0	164.8	182.1	173.6	168.0	170.6
held_cook_1	95.3	124.8	166.1	127.7	127.8	130.0
held_cook_2	113.1	191.4	206.7	201.7	204.2	204.2
held_cook_3	99.1	128.3	166.6	137.0	131.2	133.9
held_cook_4	114.1	165.8	182.4	171.1	167.2	175.3

Table 15: Identification ladder (E1) per stranger on *Asymm*, $n = 12$ strangers: mean return (fed_reeval) of the blueprint, a best response to the uniform prior at $U = 100$, the converged program-generalist, the best response to the posterior's MAP program, the full posterior of the observation phase (= TTBR), and a best response to the partner's true program. Single point estimate per stranger from the E1 evaluation tree, and this tree's own blueprint column differs by a few points from Table 7's (two different `fed_reeval` evaluation trees for the same checkpoint). Paired-gain 95% t-intervals (R1–R4) over these strangers are reported in the main text.

Stranger	TTBR	Gain
Cramped Room		
`train_sig_a`	191.9	+68.4
`train_sig_b`	207.0	+67.0
`train_cook_a`	191.9	+68.4
`train_cook_b`	191.4	+61.4
`train_flex_a`	190.9	+20.0
`train_flex_b`	200.3	+39.4
Wide		
`train_sig_a`	187.5	+358.8
`train_sig_b`	64.4	+221.2
`train_cook_a`	−12.2	−12.2
`train_cook_b`	−14.7	−14.7
`train_flex_a`	177.5	+311.2
`train_flex_b`	114.4	+213.8
Asymm		
`train_sig_a`	188.0	+76.4
`train_sig_b`	206.2	+97.2
`train_cook_a`	188.0	+76.4
`train_cook_b`	138.3	+21.7
`train_flex_a`	188.3	+85.5
`train_flex_b`	196.7	+90.2

Table 16: Second, disjoint stranger set (W7-D, defined before the method existed), blueprint 1000, final recipe: 6 strangers per layout (`fed_reeval`, 32 episodes/partner/seat). Gain is TTBR − blueprint (blueprint return itself is not separately stored in `w7d` but equals TTBR − Gain exactly). Point estimates only, with the 95% t-intervals over each six-stranger set in Table 2.

Layout	Planner	$H = 8$	$H = 16$	$H = 32$	$H = 64$
Cramped Room	Posterior-started	−0.1 [−22.7, +22.6]	+10.1 [−10.1, +30.4]	+37.6 [+16.2, +58.9]	+46.7 [+17.3, +76.2]
	Exact-program	+0.9 [−21.4, +23.1]	+12.7 [−5.5, +30.8]	+45.4 [+24.7, +66.0]	+49.0 [+21.5, +76.5]
Wide	Posterior-started	+63.7 [−10.1, +137.5]	+87.1 [+12.2, +162.0]	+99.9 [+35.5, +164.4]	+116.5 [+62.0, +171.0]
	Exact-program	+69.2 [+1.5, +136.8]	+88.7 [+14.5, +162.9]	+105.9 [+45.7, +166.2]	+123.0 [+73.5, +172.5]
Asymm	Posterior-started	−2.8 [−10.2, +4.6]	+21.8 [+18.2, +25.5]	+70.4 [+60.7, +80.2]	+63.0 [+50.8, +75.2]
	Exact-program	−0.3 [−9.9, +9.2]	+25.8 [+21.4, +30.3]	+70.7 [+59.8, +81.6]	+62.4 [+50.2, +74.7]

Table 17: Decision-time union search (G8): mean gain over the 12 common strangers with its 95% t-interval over strangers, at rollout count $N = 8$ and horizon H, on blueprint 1000. "Posterior-started" plans from TTBR's own posterior after the observation phase, and "exact-program" is handed the partner's true program.

Layout	$U = 10$	$U = 25$	$U = 50$	$U = 100$
Cramped Room	+21.0 [+10.7, +31.4] (0.31)	+36.1 [+21.9, +50.4] (0.52)	+56.7 [+32.5, +81.0] (0.82)	+68.9 [+39.4, +98.3] (1.00)
Wide	+110.4 [+42.2, +178.6] (0.66)	+118.2 [+46.6, +189.8] (0.71)	+141.5 [+45.2, +237.8] (0.84)	+167.6 [+58.9, +276.3] (1.00)
Asymm	+17.2 [+11.4, +23.0] (0.24)	+44.4 [+27.2, +61.6] (0.62)	+62.3 [+45.8, +78.8] (0.87)	+71.8 [+54.4, +89.1] (1.00)

Table 18: TTBR fine-tuning budget curve (G4): mean gain over strangers ($n = 14$ cramped, $n = 12$ wide/asymm) with its 95% t-interval at U PPO updates, and in parentheses the fraction of the $U = 100$ gain kept for that budget.

Target seed	Blueprint	TTBR	Gain
seed 2000	195.0	216.2	+21.2
seed 2001	189.7	216.4	+26.7
seed 2002	116.6	193.3	+76.7
seed 2003	192.2	211.4	+19.2
seed 2004	212.8	210.3	−2.5
seed 2005	54.7	104.8	+50.2
seed 2006	88.1	206.6	+118.4
seed 2007	154.4	217.5	+63.1

Table 19: Per-target results against the 8 held-out self-play neural strangers on *Cramped Room* (E3, final recipe, 2 fine-tune seeds per target, `fed_reeval`). Each row is a single point estimate for that target seed, and the pooled 95% t-interval over the 8 targets is reported in the main text.

Target seed	Blueprint	TTBR	Gain
seed 2000	125.3	117.3	-8.0
seed 2001	-55.9	67.2	$+123.1$
seed 2002	-94.4	-31.6	$+62.8$
seed 2003	-125.6	15.8	$+141.4$
seed 2004	145.0	136.2	-8.8
seed 2005	-88.8	15.6	$+104.4$
seed 2006	80.9	56.1	-24.8
seed 2007	-169.7	54.4	$+224.1$

Table 20: Per-target results against the 8 held-out self-play neural strangers on *Wide* (E3, final recipe, 2 fine-tune seeds per target, `fed_reeval`). Each row is a single point estimate for that target seed, and the pooled 95% t-interval over the 8 targets is reported in the main text.

Target seed	Blueprint	TTBR	Gain
seed 2000	204.1	233.0	$+28.9$
seed 2001	55.0	63.9	$+8.9$
seed 2002	163.1	223.9	$+60.8$
seed 2003	170.3	240.0	$+69.7$
seed 2004	142.8	209.4	$+66.6$
seed 2005	222.8	236.7	$+13.9$
seed 2006	154.7	213.0	$+58.3$
seed 2007	171.9	216.6	$+44.7$

Table 21: Per-target results against the 8 held-out self-play neural strangers on *Asymm* (E3, final recipe, 2 fine-tune seeds per target, `fed_reeval`). Each row is a single point estimate for that target seed, and the pooled 95% t-interval over the 8 targets is reported in the main text.

Method	Mean	Std./half-width	95% CI
published AD	-7.9	4.8	—
published DPT	-15.9	4.6	—
published Random	0.0	—	—
blueprint	-8.7	4.9	$[-11.0, -6.4]$
search (library)	17.0	3.3	$[15.5, 18.5]$
search (union grid)	17.1	2.9	$[15.7, 18.4]$
TTBR-400 (ungated = gated)	2.7	2.6	$[1.5, 4.0]$
TTBR-100 (fine-tuned at benchmark horizon)	4.6	4.3	$[2.6, 6.7]$
generalist	2.7	2.7	$[1.5, 4.0]$
prior ($U = 100$)	2.4	2.6	$[1.2, 3.6]$

Table 22: ICRL4AHT external benchmark on *test_time_wide*, under the benchmark's own protocol (100 episodes $\times$ 100 steps, greedy decoding, 20 held-out instances). The second column is the published half-width for the two in-context baselines and the standard deviation across the 20 instance means for every other row, and the last column is a 95% t-interval over instances where available (not given in the source paper for the published rows). "—" marks a value not reported by that source.

Instance	Blueprint	TTBR	Gain
assembly_line_t0	−220.0	0.0	+220.0
assembly_line_t1	8.1	46.2	+38.1
assembly_line_t2	−206.2	−0.6	+205.6
assembly_line_t3	−53.8	49.1	+102.8
assembly_line_t4	−101.9	35.3	+137.2
territory_t0	−215.6	−0.9	+214.7
territory_t1	−85.0	53.1	+138.1
territory_t2	−233.1	47.2	+280.3
territory_t3	−93.8	51.2	+145.0
territory_t4	−95.0	53.4	+148.4
utility_greedy_t0	−95.0	54.1	+149.1
utility_greedy_t1	−95.0	54.1	+149.1
utility_greedy_t2	−95.0	54.1	+149.1
utility_greedy_t3	−95.0	54.1	+149.1
utility_greedy_t4	−95.0	54.1	+149.1
recipe_aware_t0	−220.0	−1.9	+218.1
recipe_aware_t1	−221.2	−1.6	+219.7
recipe_aware_t2	−221.2	−0.3	+220.9
recipe_aware_t3	−221.2	−0.6	+220.6
recipe_aware_t4	−220.0	−1.9	+218.1

Table 23: Per-instance TTBR result on the 20 held-out ICRL4AHT instances of *test_time_wide*, blueprint 1000, final recipe, under *our* protocol (400 steps, sampled decoding, `fed_reeval`), the same instrument as the ICRL4AHT row of Table 4. The five utility-greedy instances are behaviourally identical on this layout and give identical rows. Each cell is a single point estimate per instance, and the pooled 95% t-interval over the 20 instances is reported in the main text.

Method	Cramped (14) mean [95% CI]	Wide (12) mean [95% CI]
blueprint (self-play 1000)	121.0	−77.7
FCP	75.2 [26.4, 124.1]	−18.1 [−22.1, −14.0]
dea	73.5 [60.3, 86.7]	−12.2 [−23.0, −1.5]
dea_both	71.8 [57.8, 85.7]	−10.7 [−22.5, 1.1]
dea_raw_mp	71.0 [64.4, 77.6]	−8.0 [−30.1, 14.0]
generic	75.6 [65.4, 85.7]	−12.2 [−18.1, −6.3]
none_cf	75.5 [58.5, 92.4]	−11.1 [−26.2, 3.9]
MEP $\alpha = 0.01$	55.6 [43.9, 67.3]	−8.3 [−11.5, −5.1]
MEP $\alpha = 0.05$	58.1 [55.4, 60.8]	−23.2 [−26.6, −19.8]
converged program-generalist	152.4	27.4
TTBR (final recipe)	188.2	75.8

Table 24: Train-time baselines and TTBR, mean return over the full first stranger set of each layout (14 strangers on *Cramped Room*, 12 on *Wide*), every policy scored with `fed_reeval` (32 episodes per partner per seat, eval seed 0). Population methods (FCP, the five `dea`/`generic`/`none_cf` variants, MEP at both bonus weights) are 5 ego seeds each, with the mean and 95% t-interval over the per-seed means. The blueprint, the converged program-generalist and TTBR are single policies, so no seed interval exists for them. This table replaces the earlier scoring of the same egos on the eight `held_*` strangers only, which it reproduces within one point when restricted to those eight.

Stranger	Blueprint	FCP	Variants	MEP $\alpha = 0.01$	MEP $\alpha = 0.05$	Generalist	TTBR
cook	143.1	59.4	45.2	12.8	3.9	178.4	212.2
flex	140.0	59.3	43.3	13.1	3.0	185.8	212.2
flex_conv1	138.1	59.3	38.8	11.7	4.2	180.2	212.2
signaller	112.2	0.0	81.9	−0.1	−0.1	92.5	100.2
held_sig_1	72.2	57.9	50.8	7.1	2.3	63.6	168.9
held_sig_2	167.2	142.0	92.3	150.1	171.0	202.0	210.5
held_sig_3	178.8	132.9	155.9	129.1	147.6	193.9	198.0
held_sig_4	172.2	140.2	169.4	142.4	154.4	182.9	180.2
held_cook_1	25.3	6.4	15.8	0.4	0.1	86.8	158.0
held_cook_2	147.8	139.2	87.2	152.1	172.0	202.0	210.5
held_cook_3	160.6	121.5	142.3	128.9	145.1	188.8	191.6
held_cook_4	78.4	49.5	51.9	6.4	2.2	55.3	167.7
cramped_deliver_self:a	148.4	59.8	43.5	10.8	3.1	181.5	212.2
cramped_deliver_self:b	10.0	25.7	9.9	13.6	3.9	139.4	200.3
mean	121.0	75.2	73.5	55.6	58.1	152.4	188.2

Table 25: Per-stranger return of every train-time method on *Cramped Room*, $n = 14$ strangers, the values behind Figure 2 (the re-scoring of every train-time ego on the full stranger set, `fed_reeval`, 32 episodes per partner per seat). FCP and MEP are means over 5 ego seeds, "Variants" is the mean over the five population variants (each over 5 seeds), and the blueprint, the generalist and TTBR are single policies. The rows `held_cook_1`, `held_cook_4`, `held_sig_1` and the two `cramped_deliver_self` programs are the parking strangers of RQ1.

Stranger	Blueprint	FCP	Variants	MEP $\alpha = 0.01$	MEP $\alpha = 0.05$	Generalist	TTBR
cook	0.0	−50.4	−17.1	−20.8	−27.2	−21.7	−12.2
flex	−155.6	−0.6	−6.2	2.5	−88.5	60.8	198.4
flex_conv1	−118.8	0.0	−20.9	−3.5	62.1	54.6	148.4
signaller	−145.0	0.0	−6.0	2.2	−67.9	62.9	198.4
held_sig_1	−198.8	0.0	−7.5	1.8	−60.4	59.8	115.6
held_sig_2	−63.1	0.0	−5.4	−0.8	14.1	57.7	81.2
held_sig_3	−159.4	−0.1	−6.0	−1.1	4.9	57.1	63.8
held_sig_4	−88.1	0.0	−2.9	−1.2	−11.9	44.0	154.1
held_cook_1	0.0	−54.6	−14.4	−21.1	−28.4	−12.1	−9.4
held_cook_2	0.0	−37.6	−15.6	−24.9	−24.1	−16.2	−9.7
held_cook_3	−3.1	−46.4	−16.6	−22.0	−30.4	−12.3	−10.6
held_cook_4	0.0	−27.0	−12.0	−10.4	−20.4	−5.8	−8.8
mean	−77.7	−18.1	−10.9	−8.3	−23.2	27.4	75.8

Table 26: Per-stranger return of every train-time method on *Wide*, $n = 12$ strangers, the values behind Figure 2 (the re-scoring of every train-time ego on the full stranger set, `fed_reeval`, 32 episodes per partner per seat). FCP and MEP are means over 5 ego seeds, "Variants" is the mean over the five population variants (each over 5 seeds), and the blueprint, the generalist and TTBR are single policies.

F ADDITIONAL ABLATIONS AND DIAGNOSTICS

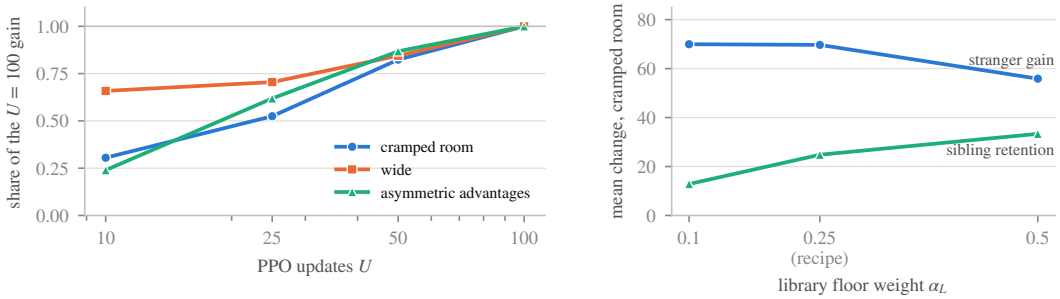

Figure 10: **Budget and floor weight.** Left: share of the $U = 100$ gain kept at smaller budgets. Right: stranger gain and sibling retention on cramped room against the library floor weight (recipe: 0.25).

Retention, floor weight and coverage. Figure 10 and Table 27 collect the ablations that fixed the recipe. *Retention.* With the self-play quarter of episodes as the only anchor the adapted agent loses 56 points with the worst held-out sibling. Raising the share to one half leaves -28 and a KL trust region -19, and only the library floor keeps it above the blueprint with every sibling ($+16.6$ to $+30.7$) at 1.02 of the stranger gain. *Floor weight.* Floors of 0.1, 0.25 and 0.5 give stranger gains of $+70.0$, $+69.7$, $+55.9$ and sibling gains of $+12.8$, $+24.0$, $+33.4$: a monotone trade, as Corollary 2 predicts, with no weight that harms a sibling. *Coverage.* The one stranger TTBR harmed on cramped room (-7.0) had a knob its programs did not vary. Adding that axis raised the posterior's fit from 0.51 to 0.99 and the gain to $+21.4$ without moving any other stranger. *Deployment check.* Keeping the blueprint whenever the simulated return against the belief never rose turns the five cook-seat failures on wide into zeros and changes nothing else (post hoc on these layouts). A gate on posterior fit would be the wrong safeguard, since fit does not predict harm across our 58 adaptation runs (below).

Table 27: **Ablations on cramped room** (14 strangers, two fine-tuning seeds, blueprint 1000). "Worst sibling" is the largest loss against a held-out sibling (mean change in brackets), and "gain kept" is the arm's mean stranger gain relative to the recipe's, or, for the anchor rows, relative to the unfloored fine-tune.

axis	arm	worst sibling (mean)	gain kept
anchor	self-play share 0.25, no floor	-56	1.00
	self-play share 0.5	-28.3	0.97
	share 0.25 + KL trust region	-19.1	0.88
	share 0.25 + library floor 0.25 (recipe)	$+16.6$ ($+24.0$)	1.02
floor weight	0.1	0.0 ($+12.8$)	1.00
	0.5	$+28.8$ ($+33.4$)	0.80
budget	$U = 10$ / 25 / 50	–	0.31 / 0.52 / 0.82
partner term	MAP program instead of the posterior	–	1.05
	true program (oracle best response)	–	1.10

Budget. With the final recipe, $U = 10$, 25 and 50 updates keep 0.31, 0.52 and 0.82 of the $U = 100$ gain on cramped room, 0.66, 0.71 and 0.84 on wide and 0.24, 0.62 and 0.87 on asymmetric advantages. With the true program in place of the posterior, $U = 25$ already keeps 0.78 on cramped room, so the posterior recipe needs about twice the budget of the oracle one. Beyond $U = 100$ we saw no further gain on the one partner we tried ($U = 200$ on the LBF partner that walks away from the food: $+0.079$, still below the $+0.10$ its pre-registered rule asked for).

Observation length. We re-ran the observation phase with $k \in \{25, 50, 100, 200, 400\}$ steps per episode and compared the resulting beliefs with the full-episode belief. On cramped room the most likely program is already the full-episode one after 50 steps for 13 of 14 strangers, and on asymmetric advantages after 25 to 50 steps for 11 of 12, and the rest of the posterior settles by 200 steps (median total variation to the full belief 0.078 and 0.097), which is the value that passes the pre-registered

stability rule on both layouts. On wide no value passes, because programs that differ only in knobs inert on the observed seat receive exactly equal likelihood and tie, and the rule counts a tie as never identified. Read post hoc as a set, the tied set is stable from 25 to 50 steps for 9 of 12 wide strangers. Figure 11 shows both quantities against the observed steps. The number of observed episodes E was not varied, so whether the 16 episodes per seat could be cut to a few, as this per-episode result suggests, is untested.

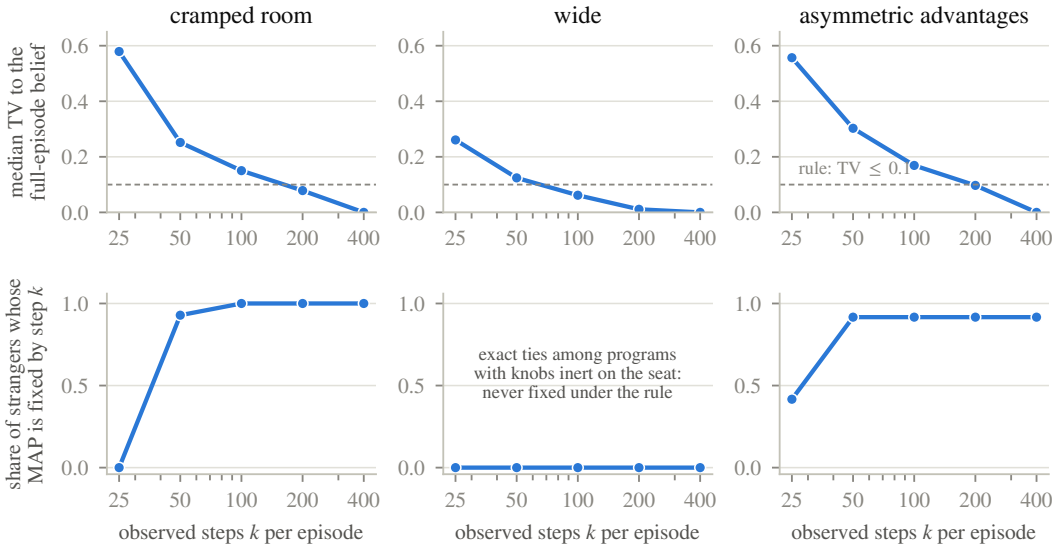

Figure 11: **Observation length.** Top: median total variation between the belief after k observed steps of each episode and the full-episode belief (the pre-registered rule asks for at most 0.1). Bottom: share of strangers whose most likely program after k steps is already the full-episode one. On wide the rule counts the exact ties among programs with knobs inert on the observed seat as never fixed.

The posterior mixture against its MAP. The posterior mixture minus its most likely program (rule E1-R3) is -3.5 $[-7.0, -0.1]$ on cramped room (the MAP marginally better), $+12.8$ $[-9.6, +35.3]$ on wide (the mixture ahead, interval including zero), -0.9 $[-3.2, +1.4]$ on asymmetric advantages and $+0.003$ on LBF, so by the pre-registered band of ± 10 the MAP suffices on three layouts and the rule is INCONCLUSIVE on wide.

Fit does not predict harm. Across the 58 adaptation items of the first blueprint (14 cramped-room, 12 wide, 12 LBF strangers and 20 ICRL4AHT instances), the realised predictive probability of the partner's actions under the belief (its fit) does not predict the gain: ICRL4AHT instances with fit 0.23 gain $+140$, a wide stranger with fit 0.64 gains $+298$, and the harmed cramped-room stranger had fit 0.51 before and 0.99 after the coverage fix. A gate on fit would remove gains and keep harms (Figure 12).

Deployment check. Keeping the blueprint when the mean return against the belief over the last quarter of the adaptation updates does not exceed that over the first quarter changes exactly the five cook-seat rows of wide (their gains become zero, and the wide mean gain moves from $+168.2$ to $+172.4$) and no other row on any layout. On the layouts reported here it was defined after those rows had been seen and is therefore a post-hoc check.

Grammar extension and model class for planning. A richer program grammar with rule switching every 50, 100 or 200 steps or at a per-step hazard, and with seat-role swapping on the asymmetric layouts (up to 1,277 hypotheses on wide), changed the planner's gains by at most ± 3 on every layout in an early, blind-evaluator run and was not pursued. With the library as its only hypothesis class the planner from TTBR's posterior gains -34.3 $[-50.9, -17.8]$ on cramped room, $+27.5$ on asymmetric advantages (against $+70.4$ with programs) and $+93.9$ on wide (against $+99.9$).

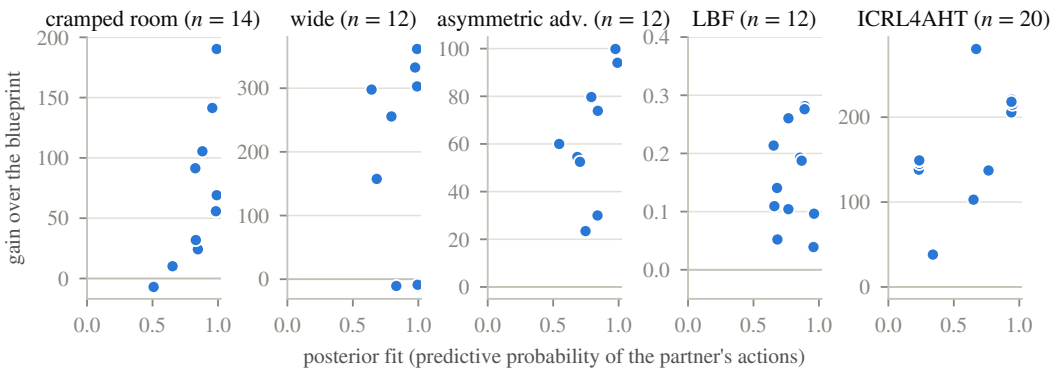

Figure 12: **Posterior fit does not predict the gain.** Each point is one adaptation item of the first blueprint: the realised predictive probability of the partner's actions under the averaged belief against the paired gain of the adapted policy.

Planner budget and information. The planner's gain is limited by its horizon rather than by its rollouts: more rollouts per action buy nothing, while lengthening the horizon from 8 to 32 steps lifts planning from TTBR's posterior from -0.1 to $+37.6$ on cramped room, $+63.7$ to $+99.9$ on wide and -2.8 to $+70.4$ on asymmetric advantages, and at 64 steps the gains flatten ($+46.7$, $+116.5$) or fall ($+63.0$). At its best horizon planning reaches 0.78 of TTBR's gain on cramped room, saturates below it on wide and closes the gap on asymmetric advantages. The observed posterior is what planning lacked: a planner that must identify the partner inside the scored episode gains $+6.0$ on cramped room, and handed TTBR's posterior it gains $+37.6$ (paired $+31.5$ $[+14.5, +48.5]$, and $+29.1$ and $+26.5$ on the other blueprints). On every layout planning from the inferred posterior does as well as planning from the exact program (0.83, 0.94 and 0.99 of the oracle planner's gain on cramped room, wide and asymmetric advantages), so posterior uncertainty is not what limits planning. Figure 13 shows the paired differences behind RQ4's first point per stranger and blueprint.

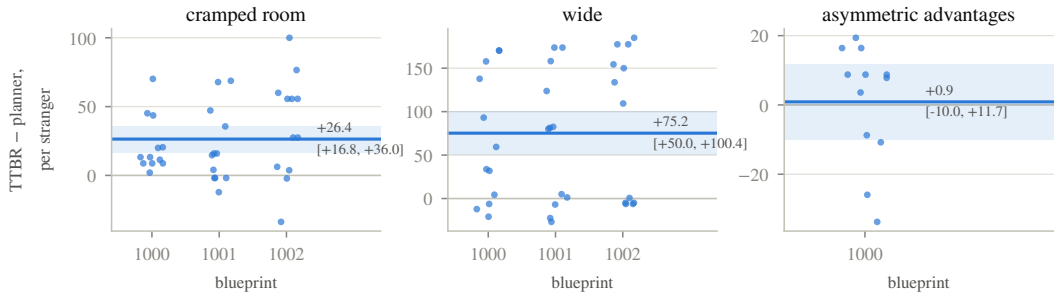

Figure 13: **Paired differences, TTBR minus the same-posterior planner at** $H = 32$, per stranger and blueprint, with the pooled mean and its 95% interval.

G FAILURE ANALYSIS: THE PASSIVE PARTNER

The five cook-seat strangers of the wide layout (`cook`, `held_cook_1` to `held_cook_4`) wait for the signaller seat to act. The unadapted blueprint on the signaller seat does not produce what they wait for, so every observation episode ends with a return of exactly zero, the partner's recorded actions are the same waiting actions under every hypothesis valid on that seat, and the posterior stays uniform over the 72 such programs (and, for the cook seat, over the library too). The fine-tunes then never start learning: their return against the belief stays flat and their gains against the true partner are -9 to -12. Three one-change arms, registered before they ran, isolate the cause. Removing the library floor, so that the posterior term gets a larger share of the environments, restores none of the five. Doubling the budget to $U = 200$ restores none. Replacing the posterior by the partner's true program, with the floor kept, restores two of the five, and the true-program recipe without the floor had earlier

reached $+27$ to $+165$ on all five. Identification from a passive episode is therefore the cause, and the floor's smaller scripted share slows learning further once the belief is uninformative. Planning on top of the adapted policy repairs the seat ($+48.9$ over the adapted policy on wide, from about -12 to $+31$ to $+71$ on these five), because a lookahead that samples cooperative behaviour sees the partner respond. An observation phase that does the same is the natural repair on the identification side.

H LIMITATIONS IN DETAIL

The passive partner. The method's one systematic failure is the cook seat of the wide layout. Five strangers there do nothing until the ego cooperates, so an observation phase played by a blueprint that does not cooperate ends with an episode return of exactly zero and a posterior uniform over the 72 programs valid on that seat: the regime $D_N = 0$ of Theorem 1, in which no later compute can help. The fine-tunes never start learning (-9 to -12) although the same recipe with the partner's true program reaches $+27$ to $+165$ on the same five partners (three one-change repairs in Appendix G). Planning on top of the adapted policy repairs it (RQ4), because its lookahead sees what the observation did not. An observation phase that elicits distinguishing behaviour is the natural next step.

The assumptions. A1 grants a simulator at test time, as decision-time search does, and the observation phase records the states at which the partner acted, which the program likelihoods and the replay of library policies need. A2 is the load-bearing one: our strangers are drawn from the same family as the program class, off its grid but inside its span. The one stranger the method harmed had a live knob the grid did not vary (Appendix F), and outside the family, on the ICRL4AHT teammates, search is better and a generalist ties (RQ5). Neural strangers sit in between, and there most of the gain is generic rather than partner-specific. A3 and the cost are real: 16 observed episodes per seat and 3 to 7 million simulator steps, or 40 to 70 minutes of eight CPU cores, before the improved policy earns anything back. The number of observed episodes was never varied, although the per-episode study of Appendix F, in which the most likely program is fixed within 25 to 50 steps of an episode, suggests that far fewer would do. A4, stationarity, was never tested: every partner is a fixed policy in all three phases, and the harm table of RQ1, whose worst pairs cost 83 points on cramped room and 121 on wide, measures what an adapted policy is exposed to if the partner turns into another member of the family after adaptation.

Rivals not run. A library of best responses precomputed at train time for the fixed grid of 72 to 288 programs, or for its distinct behaviours, and selected at test time by the same posterior would cost nothing per partner, and since the MAP program suffices in our data (RQ3) it would plausibly match TTBR in-family. What it cannot do is respond to a partner off the grid, to a class too large to precompute, or to the neural strangers of RQ5, and we did not measure how much of TTBR's gain those cases carry. RQ4 therefore establishes that the compute is better spent on improvement than on planning, not than on selection.

Statistical scope and the claims we do not make. The strangers hold 10 of 14 distinct behaviours on cramped room, 11 of 12 on wide, 9 of 12 on asymmetric advantages and 12 of 12 on LBF, so intervals over strangers are optimistic, and Table 2 gives the intervals over distinct behaviours beside them. The intervals of the second and third asymmetric-advantages blueprints were computed from the stored per-stranger gains after the fact and were not registered. On wide the planning comparison at $H = 64$ has an interval that includes zero, and the neural-stranger gain, though positive, spans $+3.5$ to $+150.1$. We do not claim that policy improvement beats planning in general (it ties on asymmetric advantages and loses outside the family), that planning is brittle under posterior uncertainty (planning from the inferred posterior matched planning from the exact program), that the posterior mixture matters (its MAP does as well on three layouts and trails it by 12.8, with an interval including zero, on wide), or that identification is needed everywhere (a generalist is better on asymmetric advantages). Our MEP baseline simplifies the ego stage (§4.2), and no human partners were run.

I PRE-REGISTRATION RECORD AND REPRODUCIBILITY

Rules and verdicts. Every experiment was run under a written decision rule fixed before its jobs were submitted, with thresholds that were never changed after data arrived. Tables 28 and 29 list the

rules that the main text relies on and their verdicts, including the ones that did not go the paper's way. Statistics computed after a result had been seen (the set-level reading of the observation-length rule on wide, the deployment check on the layouts where it was defined) are identified as post hoc wherever they appear. Two items were added after the record had been closed: the generalist-start arm on the two control layouts, whose outcome (no addition, by Corollary 1) was registered before it ran and which is reported whatever it read, and a re-scoring of the train-time baselines on the full first stranger set with the paper's instrument (they had been scored on the eight `held_*` strangers only), reported as it fell. The dated rule files, with their commit timestamps, will be released with the code so that the record can be checked.

Table 28: Pre-registered rules and their verdicts, part 1 of 2 (the wave-1 and wave-2 rules and the identification ladder). Rule texts are transcribed from the dated rule files. Thresholds are in return units, "ahead" means a positive paired difference, and a fine-tune's *sibling retention* is its mean change against the four held-out siblings.

item	rule	verdict
G1, reachability	parkers gain $\geq +40$ on 3 of 5; wanderers $\geq$ oracle search; the mean over strangers of the fine-tunes' sibling retention ≥ -10	PASS
A2.1, posterior version	parkers keep ≥ 0.5 of G1's gain; worst stranger ≥ -5; the siblings' mean retention ≥ -10	INCONCLUSIVE by rule: 0.92 of the gain kept, worst stranger -44.8, worst fine-tune's sibling retention -56.4
A2.3, anchors	an arm keeps every fine-tune's sibling retention ≥ -10 and ≥ 0.8 of G1's gain	INCONCLUSIVE: neither arm (self-play share 0.5 worst -28, KL -19)
A2.8, library floor	an arm keeps every one of the 14 fine-tunes' sibling retention ≥ -10 and ≥ 0.8 of A2.1's gain	PASS ($+16.6$ to $+30.7$, 1.02 of the gain)
Final recipe per layout	cramped: parkers keep ≥ 0.5 of G1's gain, every fine-tune's sibling retention ≥ -10, worst stranger ≥ -5; wide: 2 of 3 parkers at $+40$, siblings' mean retention ≥ -10; ICRL: interval above 0, worst instance ≥ -5; LBF: ≥ 0.5 of G1's gain	PASS on all layouts except one cramped stranger at -7.0 (fixed by the coverage arm)
E2, blueprints	each new blueprint $\geq 0.5\times$ blueprint 1000's gain, with its layout's retention and worst-stranger conditions	REPLICATES on cramped, wide, LBF; asymm 1001 REPLICATES, 1002 NOT ($+29.4 < 0.5 \times 71.3$)
E1-R1, generalist	TTBR $-$ generalist $\geq +20$, interval above 0	TEST-TIME NEEDED on cramped ($+35.8$), wide ($+48.4$); GENERALIST BETTER on asymm (-11.4); suffices on LBF
E1-R2, prior	TTBR $-$ prior best response	$+83.6$, $+70.4$; prior suffices on asymm, LBF
E1-R3, MAP	$\lvert$mixture $-$ MAP$\rvert < 10$	MAP SUFFICES on cramped, asymm, LBF; INCONCLUSIVE on wide (mixture ahead by $+12.8$, interval includes 0)
E3, neural strangers	mean gain interval above 0, 6 of 8 positive	PASS cramped, asymm; INCONCLUSIVE wide
E4/G2, ICRL4AHT	report only	search $+17.0$, TTBR $+2.7$ / $+4.6$
G1/G8, horizon	planning ≥ 0.8 of TTBR's gain at some horizon	PARTIAL cramped (0.78); SATURATES wide; CLOSES GAP asymm

Table 29: Pre-registered rules and their verdicts, part 2 of 2 (the planning comparison and wave 7). Conventions as in Table 28.

item	rule	verdict
G5, same posterior	planner from TTBR's posterior ≥ 0.8 of exact-program planner	RECOVERS ORACLE cramped, wide; MATCHES asymm (the brittleness sentence not licensed)
G8, paired at $H = 64$	TTBR − planner, interval above 0	$+12.9$ $[+7.0, +18.8]$ cramped; $+51.7$ $[-3.2, +106.7]$ wide
W7-B, three blueprints	pooled interval above 0 and ≥ 9 of 12 ahead per blueprint	PARTIAL on both (pooled intervals above 0; one blueprint per layout at 8 of 12)
G9, from the generalist	adds $\geq +20$; ends within ± 10 of TTBR from self-play	COMPOSES, EQUAL START on cramped room and wide
G9 on the control layouts (run last; predicted NO ADDITION)	as G9, LBF thresholds scaled by $1/400$	NO ADDITION on both: asymm. $+6.6$ $[+3.9, +9.3]$, BETTER START $(+17.2)$; LBF 0.000, EQUAL START
Train-time baselines re-scored on the full first set (run last; no rule)	report only; ordering of Table 3 expected to hold	holds: FCP 75, variants 71–76, MEP 58 against the blueprint's 121 on cramped room; -18, $-12 \ldots -8$, -8 against -78 on wide
G10, observation length	belief within TV 0.1 of the full belief and same MAP for ≥ 11 of 12	200 steps PASS on cramped, asymm; no value on wide (exact ties)
G11, planning on TTBR	adds $\geq +20$	ABSORBED cramped (-3.6); COMPOSES wide $(+48.9)$, asymm $(+24.1)$
G12, MEP	MEP within 20 of TTBR	MEP BELOW on both layouts
W7-A, conflict	$\geq 25\%$ conflicting pairs on cramped; no single fitting response on wide; 0% on asymm, LBF; ordering holds	ordering HOLDS; strict form FAILED (21%; one wide response fits)
W7-C, asymm blueprints	as E2	1001 REPLICATES $(+40.5)$; 1002 NOT $(+29.4)$
W7-D, second set	mean gain interval above 0, 5 of 6 positive	PASS cramped, asymm; INCONCLUSIVE wide
W7-E, library-only planner	report only	yes on cramped, asymm; no on wide
W7-F, floor weight	0.5 arm keeps ≥ 0.8 of the recipe's gain; no loss with any sibling	yes

The evaluator defect and its repair. Until the fourth day of the study the search evaluator fed every trained network zeros on one input channel, the teammate's previous action, while the trainer and its evaluator fed the true value. Training was never affected, so every trained policy is valid, but every number produced by the search evaluator before the repair described policies that were blind on that channel. The defect was found when the self-play reference on the wide layout scored -63 with a copy of itself instead of $+222$. Every policy in this paper was re-scored with the repaired instruments, and every pre-registered verdict taken before the repair survived re-scoring, while absolute numbers moved (by about 10% on cramped room and LBF, and by hundreds of points on wide, where the channel carries the signalling convention). The old files are reproducible byte for byte with the flag that zeroes the channel, and no number measured before the repair is quoted in this paper. One check was run on the negative result that motivated the library floor: the no-floor fine-tunes (three quarters of the adaptation episodes against the belief, one quarter self-play) were retrained from the corrected beliefs, 44 of 46 of which kept the same most likely program, on the 14 cramped-room and 12 LBF strangers with two seeds each. The forgetting reproduced, with a worst change against a held-out self-play partner of -46.8 against -56.4 before the repair and 0.90 of the true-program gain kept on the parking strangers against 0.92, so the forgetting comes from the recipe and not from the defect.

J Earlier lines and negative results

The study that produced TTBR closed several lines before it, each under a written rule, and we list them so that the reader can see what did not work. Their search-based numbers were measured with the defective evaluator at the default horizon and are not quoted. (1) *Partner-space flatness*: merging a population into one policy by a sharpness-aware objective over partners won 0, tied 4 and lost 3 comparisons against one ego. (2) *Decision-effect representations*: representing the partner by its effect on the ego's joint action values tied every variant once all trained on the same pool. (3) *Convention discovery*: searching for conventions as game symmetries and identifying them by meta-learning found no conventions on the real layouts. (4) *Worst-case verification* over a finite partner set accepted 1 to 2% of policies with a third of them false accepts. (5) *Distilling belief-space search into a policy* hurt with a belief input and kept 0.61 of the search's gain on wide only. (6) *Liveness*: a randomised yield rule removed the stalls strangers cause (0.65 to 0.03 of episodes) without changing returns, so the stranger problem is not a deadlock problem. (7) *Solo-anchored self-play* raised the solo floor from 47 to 175 but produced an ego that was worse with strangers. (8) *Mutual adaptation*: two adaptive planners did not collapse on the real layouts (210 to 213 against 217 in self-play), so the collapse seen in a synthetic matching game is not a benchmark phenomenon. (9) *Few-shot refinement of the specific partner* inside the search captured 0.27 of the headroom at 16 episodes. (10) *Certified adaptation*: no anytime-valid gate removed stranger harm without removing the gain. (11) *A four-term decomposition* of the coordination gap computed with the default-horizon search attributed nothing to identification, and the long-horizon results of RQ4 show this to be an artefact of the horizon, and the decomposition is not used. (12) The prediction that layout contention predicts the unattainable part of the gap across layouts failed (Spearman $+0.14$ against a registered $+0.8$).

K The use of large language models

Large language models were used as coding and writing assistants throughout this project: to implement and test the experiment code under the authors' specifications, to run and monitor the cluster jobs, to draft the analysis scripts that read the pre-registered rules, and to draft and edit the text of this paper. Every decision rule was written by the authors before the corresponding jobs ran, every number in the paper comes from the analysis scripts described in Appendix I, and the authors reviewed all text and take responsibility for it.