510K: WHEN-TO-COOPERATE UNDER LATENT TEAM RELATIONSHIPS AND THE OTHER-REGARDING VALUE OF INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents acting with others must first decide *whether and with whom* to cooperate, often by inferring relationships from observed behaviour. We study this when-to-cooperate problem in a setting where the relationship is learnable in principle yet unused in practice. We introduce **510K**, a four-player card game in which team membership is hidden at deal time and the reward itself depends on the unknown relationship. We separate four questions that benchmarks conflate — availability, decodability, utilization, and value — and report a consistent pattern. *(1) The relationship is decodable but unused.* 67% of decisions are logically determined by public play and a history probe reaches AUC $0.72/1.00$, yet 1M-step PPO never conditions on the team: actions change in only 0.5–3.6% of decisions when the team bits are zeroed, and adding more evidence, explicit revelation, stronger win incentives, or an undiscounted objective changes nothing. *(2) Information value is other-regarding.* Informing the agent is inert; informing *partners* moves outcomes — team-aware partners raise the agent's first-place rate by $+0.066$ ($p = 0.007$, replicated on a fresh deal set), adding team-awareness to both teams lowers the agent's reward relative to the ignorant condition, and training against aware bots yields behavioural adaptation without reciprocity. Small language models reproduce the same gap, and a full-information search (first-place 0.43–0.48 vs. the best policy's 0.21–0.36) shows large headroom. We release the environment, the evaluation protocol, and belief-labelled trajectories as an instrument for when-to-cooperate decisions.

1 INTRODUCTION

Multi-agent benchmarks usually hand the agent its social structure: teams are fixed and known, and only the question of *how* to cooperate remains. Realistic settings instead require deciding *whether and with whom* to cooperate from partial evidence. In SMAC (Samvelyan et al., 2019) and the StarCraft family teams are known; Hanabi is fully cooperative; Overcooked studies ad-hoc teamwork with a known partner; Dou Dizhu has a public landlord role. Hidden-role games (Avalon, Werewolf, The Resistance) do involve identity inference, but they are adversarial, communication-heavy, and mostly studied with LLMs rather than as RL testbeds with ground truth. The result is a gap: there is no compact, trainable environment in which *cooperative* team structure is latent, controllable, and measurable at the decision level — and, more importantly, no clean empirical statement about whether current agents can actually use such information.

We present **510K**, a four-player shedding card game that fills this gap as an *instrument* for a concrete, open difficulty: when-to-cooperate under latent relationships. Team membership is determined at deal time: players holding a red ace form the "red" team; the rest form the other team. A deal is 2v2 (76.5%) or 1v3/3v1 (23.5%, closed-form hypergeometric), so the latent variable includes *whether the agent has a partner at all*. The relation is hidden, so the reward function itself is indexed by a latent variable, and the same observation can map to different returns. The environment ships with an *intervention stack*: four modes (SINGLE/STATIC/DYNAMIC/OBVIOUS), a graded reveal dial that shows the team bits with probability p per decision, public scoring information (obs-v2),

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.

scripted bot policies with configurable team-awareness, and per-decision ground-truth belief labels in a replayable trajectory sample.

We use this stack to ask four questions that are usually conflated: **(Q1) availability** — is information about the team present? **(Q2) decodability** — can it be recovered from the agent's own observation and public history? **(Q3) utilization** — does the trained policy's *behaviour* depend on it? **(Q4) value** — does providing it change terminal outcomes? Our results separate these layers cleanly, and the answer pattern is asymmetric and, we argue, interesting:

1. **The relationship is decodable but unused.** 67% of DYNAMIC decisions are logically determined by public information, and a linear probe over non-team-bit features reaches AUC 0.63/0.72 (single step/history; 0.90/1.00 in OBVIOUS). Yet after 1M steps the policy shows no practically meaningful conditioning on the team: zeroing the team bits changes its argmax action in only 0.5–3.6% of decisions, and the deferral asymmetry is ≈ 0 in DYNAMIC and *negative* in OBVIOUS. Providing more evidence (public scores), explicit revelation, a human-like zero-sum settlement, or an undiscounted objective leaves all outcomes unchanged.
2. **Information value is other-regarding.** Making the information available to the *learner* is inert; making the same information available to *partners* moves outcomes. When the agent's teammates act on the hidden team, the agent's first-place rate rises by $+0.066$ ($p = 0.007$, paired over seeds; replicated on a second deal set), while the agent's own conditioning does not change. Adding awareness to the opponents *as well* (all-aware, relative to the fully ignorant condition) keeps first-place flat but lowers the agent's reward (69.2 vs. 78.2); the clean opponent-only contrast is all-aware vs. teammate-aware, which also lowers reward (69.2 vs. 80.8). Training against team-aware bots improves placement, but through a behavioural adjustment (larger combinations, fewer tricks contested), not conditional cooperation — "adaptation without reciprocity." The value of the information is therefore realized almost entirely through other players' behaviour.
3. **The problem is open, and the failure generalizes.** Small language models reproduce the same asymmetry (they read explicitly revealed teams perfectly but neither infer nor exploit hidden ones), and a full-information 1-ply search reaches first-place 0.43–0.48 versus the best policy's 0.21–0.36: the benchmark is far from saturated.

Contributions.

1. **A finding: utilization is the bottleneck, not evidence.** We separate availability/decodability from utilization/value and show that in this task current PPO neither conditions on a hidden relationship nor benefits from making it explicit — a measurable information-utilization failure under controlled interventions.
2. **A finding: information value is other-regarding.** A who-knows decomposition shows that the same information helps the agent only when *other players* use it; the agent-side channel is null. Training in such a population improves placement via behavioural adaptation, not reciprocity.
3. **A controlled instrument.** A tested Gymnasium + PettingZoo AEC environment (action masking, 86 tests) with latent deal-time teams, a graded reveal dial, observation/settlement/discount ablations, bot team-awareness variants, and per-decision belief labels in a replayable trajectory sample.
4. **Calibration and headroom.** Outcome metrics with same-seat bot references and paired same-deal comparisons, a robustness check on a second deal set, and a full-information search reference showing the task is open.

2 Related Work

2.1 Card-game RL toolkits

RLCard (Zha et al., 2020) provides many card games behind a common interface; each included game isolates a subset of challenges, and none exposes a hidden, deal-time-determined team. OpenSpiel (Lanctot et al., 2019) offers game-theoretic tooling over 50+ games, oriented toward solving

Table 1: Positioning: ● strong/present, ◎ partial, ○ absent.

Challenge	510K	DouDizhu	Hanabi	Mahjong	Tichu	Overcooked	SMAC
Immediate scoring	●	○	○	●	●	○	○
Terminal competition	●	●	○	●	●	○	●
Partial observability	●	●	●	●	◎	○	◎
Team cooperation	$\bullet_{\text{hidden}}$	$\bullet_{\text{fixed}}$	$\bullet_{\text{all}}$	○	$\bullet_{\text{fixed}}$	●	◎
Hidden/dynamic teams	●	○	○	○	○	○	○
Controlled ablation	●	○	○	○	○	◎	◎

extensive-form games rather than RL under sparse rewards. PyTAG (Balla et al., 2023) covers 20+ tabletop games with a modern API but shallow per-game depth.

2.2 Shedding-type and scoring card games

DouZero (Zha et al., 2021) established that a shedding game with sparse terminal rewards is a serious benchmark, but its landlord role is public and fixed. Tichu, Hearts and Spades have scoring and partnerships, but the partnerships are known, so the relationship itself is never inferred. Suphx (Li et al., 2020) shows expert play in Mahjong, but with fixed, known seats and no hidden team. 510K keeps the long-horizon shedding structure and adds a *hidden* relationship whose reward value the agent must infer.

2.3 Cooperative and imperfect-information benchmarks

Hanabi (Bard et al., 2020) is the canonical cooperative imperfect-information benchmark, but every agent is a teammate and there is no competition or scoring trade-off. Overcooked (Carroll et al., 2019) probes ad-hoc coordination with a *known* partner. Melting Pot (Agapiou et al., 2022) studies mixed motives over configurable social situations, but team identity is not a controllable, ground-truthed latent variable. Hidden-role games are adversarial and communication-centric; RL work such as DeepRole exists, but there is no standardized RL benchmark with per-decision ground truth.

2.4 Multi-objective and hidden-role settings

Multi-objective RL tackles trade-offs between conflicting objectives (Hayes et al., 2022); 510K is an instance with exactly two such objectives (immediate score and terminal finish) whose tension is measurable. Diplomacy/Cicero (Meta FAIR Diplomacy Team et al., 2022) and hidden-role LLM studies (Avalon, Werewolf) demonstrate strong interest in relationship inference at a much larger scale, but lack a lightweight RL testbed with ground-truth team assignments. 510K fills that gap: the latent relationship is known to the simulator, so "did the agent infer it?" is directly measurable.

2.5 Prior results on 510K

Prior work (Anonymous, 2027) studies a *phenomenon* on this environment: under hidden team assignment, policy-gradient updates from different relationships cancel in expectation, producing *deceptive stability*. The present paper is deliberately complementary: prior work asks *why* a learner stalls; we characterize the *environment* and provide the controlled ablations and behavioural probes that make the phenomenon reproducible and the mechanism studyable.

2.6 Positioning

Table 1 compares 510K with the environments above. Only 510K couples immediate scoring, terminal competition, partial observability, *hidden dynamic teams*, and controllable ablation in one game. The claim is about this combination and its controllability, not about being the hardest card game; our contribution is to make the social relationship a latent variable to be inferred rather than an input, and to make that inference failure measurable.

3 THE 510K ENVIRONMENT

3.1 RULES

Four players, 52 cards, 13 each. On a turn a player plays a combination that beats the current trick or passes; when the others pass, the trick ends and its accumulated points (5=5, 10=10, K=10) go to the trick leader, who opens the next trick. The game ends (SINGLE) when one player empties their hand, or (team modes) when both members of a team finish. There are 16 playable combination types, including bombs and the eponymous 5-10-K.

3.2 TEAMS AND REWARD

STATIC: fixed teams $\{0, 2\}$ vs $\{1, 3\}$, known. DYNAMIC: a player holding a red ace belongs to the red team; membership is hidden. OBVIOUS: the same deal-time teams, but a 4-bit red-membership mask is in the observation. The red team has size 2 with probability $1 - 4\binom{50}{11}/\binom{52}{13} \approx 0.765$ and size 1 otherwise (verified exactly), so the latent variable includes *whether the agent has a teammate at all*.

The reward is terminal only: individual$[i]$ $=$ $510K$ points $+$ (first-place bonus 15) $-$ (unfinished/last penalty 15); team_total$[t]$ sums the members of team t; and reward$[i]$ = team_total$\times$ (2 if the team wins else 1), where the winning team is the one with the higher total. We treat this scalarisation as part of the specification and study settlement variants separately.

3.3 OBSERVATION AND INTERVENTION STACK

The canonical observation (obs-v2) contains the agent's hand, the last trick, hand sizes, pass count, its own score, the four public 510K scores, and the pending trick score (117 dims; 121 in OBVIOUS with the 4 team bits). *RevealEnv* includes the team bits with probability p per decision; the mutual information of this channel is $MI(p) = p\,H(T) - \delta(p)$ and reaches at least 91% of the ideal $p\,H(T)$, so the dial is a nearly-lossless information channel. Bot policies can be ignorant (the default rule bot), teammate-aware (only the agent's teammates act on the true team), or all-aware. Terminal settlement can be the default, a zero-sum rule (winners $+$diff, losers $-$diff), or binary (± 1). The environment is distributed with the Gymnasium single-agent interface (`510K-v0`) and a PettingZoo AEC interface (Towers et al., 2024; Terry et al., 2021); action masking is built in (the pass action is masked when passing is illegal), and all 86 tests pass.

4 PROTOCOL AND RESEARCH QUESTIONS

We evaluate policies as player 0 against fixed bots on a fixed deal set (seeds 1000–1099). Because the policy and the bots are deterministic, each seed is a fixed game. Primary metrics are *first-place rate* (the agent finishes first), *team-win rate* (the agent's team finishes first — a finish-based outcome, reported separately from the reward whose winner is score-based), and *mean terminal reward*. Belief probes report the fraction of decisions where the relation is logically determined by public information and the AUC of a linear probe over non-team-bit features. Behavioural probes report the deferral asymmetry $\mathrm{asym} = P(\text{pass} \mid \text{true partner leads}) - P(\text{pass} \mid \text{opponent leads})$ and a counterfactual bit-flip rate: at each decision, zeroing or shuffling the team bits and asking whether the policy's argmax action changes.

Canonical training: MaskablePPO (Raffin et al., 2021; Schulman et al., 2017), 1M steps, 5 seeds, MLP 256×256, $\gamma = 0.99$, rule-bot opponents. We report ablations for the observation (obs-v1/v2), settlement (current/zero-sum), discount ($\gamma = 0.99/1.0$), reveal dose p, and bot team-awareness. All comparisons are paired over the same seeds or the same deals; a second deal set (seeds 2000–2099) checks deal-set robustness.

Table 2: Belief-inference probe (non-team-bit features; 1200 games/mode, 13.9k/17.8k agent decisions).

mode	logically determined	obs-probe AUC	hist-probe AUC
DYNAMIC	0.671	0.632	0.722
OBVIOUS	0.592	0.896	0.998

Table 3: Team-bit counterfactual on obs-v1 checkpoints: fraction of decisions where zeroing/shuffling the bits changes the argmax action (mean over 5 seeds).

checkpoint	decisions w/ bits	flip (zero)	KL(true‖zero)
OBVIOUS	1780	0.005	0.0086
RevealEnv p=1	1818	0.036	0.0284

5 RESULTS

5.1 THE RELATIONSHIP IS DECODABLE FROM PUBLIC PLAY (Q1/Q2)

The relation is present in the observation and history: 67% of DYNAMIC decisions are logically determined by the agent's own red aces plus publicly revealed red aces, and a linear probe over non-team-bit features reaches AUC 0.632 from a single observation and 0.722 with history (OBVIOUS: $0.896/0.998$). Adding public scores to the observation (obs-v2) changes these numbers by less than 0.01, so the agent's own cards and the revealed red aces are the main evidence. The reveal dial carries $\geq 91\%$ of the ideal $p\,H(T)$.

5.2 POLICIES DO NOT USE THE RELATIONSHIP (Q3)

Counterfactual bit interventions on revealed-team checkpoints (obs-v1) zero or shuffle the team bits at every decision: the argmax action changes in only 0.5% of decisions for the OBVIOUS checkpoint (KL 0.009) and 3.6% for the RevealEnv-p=1 variant (KL 0.028). The deferral asymmetry behaves the same way: STATIC (fixed, known partner) is strongly positive ($+0.391$), while DYNAMIC is ≈ 0 ($+0.009$) and OBVIOUS is *negative* (-0.039). Neither evidence provided by the environment nor experience with cooperative partners produces conditional cooperation on this channel.

5.3 INFORMATION VALUE IS OTHER-REGARDING (Q4)

Holding the agent policy fixed and varying only the *bots'* team-awareness decomposes who must know the team for it to matter. The canonical (rule-trained) policy in DYNAMIC:

Making the agent's *teammates* aware — the only clean causal contrast, since the agent's own information is untouched — raises the agent's first-place rate from 0.212 to 0.278 ($+0.066$, paired $p = 0.0069$) while the agent's own deferral asymmetry does not increase. Making the *opponents* aware as well keeps first-place flat but reduces the agent's reward (69.2 vs. 78.2 under fully ignorant bots). The value of the information is therefore realized almost entirely through other players' behaviour.

The pattern repeats with trained agents. A teammate-aware-trained policy gains $+0.064$ first-place from partner cooperation ($p = 0.011$) without reciprocating; an all-aware-trained policy gains $+0.094$ ($p < 0.0001$) while its asymmetry becomes more negative.

Behaviourally, the all-aware-trained policy plays larger combinations (1.15 vs. 1.11 cards per play) and contests fewer tricks (1.80–1.93 vs. 1.97–2.06): it adapts by racing to shed, not by cooperating. A second deal set reproduces the partner-cooperation effect ($+0.024$–$+0.036$, $p < 0.04$) and confirms that the all-aware advantage is opponent-conditional.

Table 4: Deferral asymmetry on the canonical (obs-v2) checkpoints; p-values are one-sample t-tests against zero ($n = 5$).

mode	asym	p
STATIC	$+0.391$	$< 10^{-4}$
DYNAMIC	$+0.009$	0.012
OBVIOUS	-0.039	$< 10^{-4}$

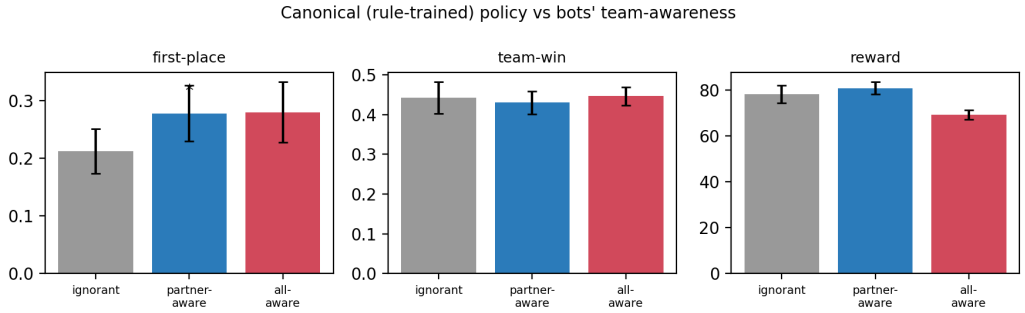


Figure 1: Who-knows decomposition: the agent policy is fixed, and only the bots' team-awareness varies. Partner-awareness raises first-place; adding awareness to both teams lowers reward.

5.4 Robustness of the nulls

Adding public scoring information (obs-v2 vs. obs-v1) changes no outcome (all paired $p > 0.14$) and no conditioning (DYNAMIC asym $+0.009$ vs. $+0.003$). Replacing the settlement with a zero-sum rule gives $+0.020$ first-place ($p = 0.61$). The controlled reveal dose–response is flat (per-seed first-place slope -0.006, $p = 0.83$; team-win $+0.017$, $p = 0.47$). Undiscounted training (γ=1.0) is not significantly different (first 0.178 vs. 0.206, $p = 0.28$). Two-versus-three team sizes shift composition-dependent metrics (first 0.200 vs. 0.221; finish-based team-win 0.503 vs. 0.314) but do not change the conclusions.

5.5 A search reference shows large headroom

A full-information 1-ply rollout reference — enumerate each legal action, finish the game with the rule-bot continuation, take the best by the chosen objective — reaches first-place 0.430–0.480 under a first-place objective (policy: 0.206–0.362) and team-win 0.66–0.77 / reward 144–162 under a reward objective (policy: 0.43–0.48 / 69–79). The task is far from saturated. The reference uses privileged information and is therefore a *search reference*, not a game-theoretic optimum.

5.6 Language models reproduce the gap

Small open models (qwen2.5-7B/14B, llama3.1-8B, 4-bit, temperature 0, one call per decision, text obs-v2) show the same asymmetry as the RL policies. Hidden-team belief is far below a trivial majority baseline (exact match 0.04–0.10 vs. 0.322 for always-guessing the most common teammate); with the true team bits in the prompt, belief is nearly perfect (0.95–1.00), yet zeroing the bits changes only 12–14% of actions. Forced-play and chain-of-thought prompts improve play (llama8 first $0.13 \rightarrow 0.19$) but not belief. Play is weak and passive overall (qwen passes 85% of decisions; llama8 reaches first-place 0.13, below the RL policy's 0.206 and the search reference's 0.43+). These results are exploratory: one prompt format per condition, small models, a local server, and no counterfactual outcome rollouts; we report them to show the failure generalizes beyond PPO.

Table 5: Who-knows decomposition: agent policy fixed, bots' team-awareness varies (5 seeds, 100 episodes/seed).

bots	first	team-win	reward	asym
ignorant	0.212	0.442	78.2	+0.002
teammate-aware	0.278	0.430	80.8	−0.012
all-aware	0.280	0.446	69.2	−0.052

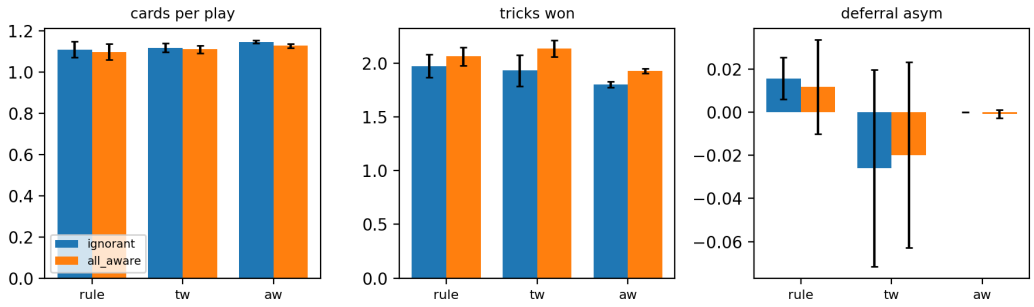


Figure 2: Behavioural profile by training population (rule/tw/aw) under ignorant vs. all-aware bots: larger combinations, fewer tricks won, and more negative deferral asymmetry as the trained population becomes more team-aware.

6 Discussion

The consistent picture is that *information value is other-regarding*. Making team information available to the learner — more evidence, explicit bits, stronger win incentives, an undiscounted objective — does not change its behaviour or its outcomes. Making the same information available to *partners* does: teammates who act on the relation raise the agent's placement, and adding team-awareness to both teams lowers the agent's reward relative to the ignorant condition. A learner trained in such a population improves its placement, but through a behavioural adjustment (racing to shed) rather than conditional cooperation.

This is a sharper statement than "agents fail to infer": the relation is available and decodable, yet the policy-to-payoff link runs through other agents, not through the learner's own conditioning. Benchmarks that report only team-level win rates can miss this structure; outcome metrics computed on the same deals, counterfactual bit interventions, and a who-knows decomposition make it visible. For ad-hoc teamwork and zero-shot coordination, the results suggest that improving *partners'* responsiveness — and the agent's ability to exploit a cooperative population — may matter more than giving the agent a better belief.

7 Limitations

- Partners and opponents are deterministic scripted policies; they are honest (no deception) and cannot reciprocate adaptively.
- The search reference has privileged information; no game-theoretic ceiling or self-play solution is established.
- The canonical evaluation uses 100 deals; absolute rates move by ~ 0.1 between deal sets (0.212 vs. 0.314 for the same policy), so we rely on paired, relative effects and report a second deal set.
- The deferral asymmetry is a raw pass-rate difference that does not control for game phase, leader identity, or hand strength; it is a behavioural channel, not a proof of internal representation.
- STATIC checkpoints are near seed-identical, so those cells have a reduced effective sample size.

Table 6: Controlled reveal dial (v1 obs, MLP, 5 seeds): no dose–response.

reveal p	first	team-win	reward
0.00	0.212 ± 0.054	0.442 ± 0.056	78.2 ± 5.3
0.25	0.184 ± 0.053	0.440 ± 0.034	79.1 ± 4.9
0.50	0.162 ± 0.024	0.428 ± 0.025	76.6 ± 3.8
0.75	0.192 ± 0.022	0.446 ± 0.052	78.1 ± 3.6
1.00	0.200 ± 0.057	0.460 ± 0.043	82.2 ± 7.6

- LLM results use one prompt format per condition and no counterfactual outcome rollouts; they are exploratory.
- No human study; a CLI with trajectory recording is provided for future work.

8 Conclusion

We separate four questions that latent-relationship benchmarks conflate — availability, decodability, utilization, and value — and report a consistent, replicable pattern in a controlled card game: the relationship is decodable from public play but unused by 1M-step PPO, and the value of the information is realized through other players, not through the learner's own conditioning. A full-information search shows the problem is far from solved, and small language models reproduce the same gap. We release the environment, protocol, and belief-labelled trajectories as an instrument for studying when-to-cooperate decisions under coupled objectives.

Use of AI Tools

In this work, we used generative AI tools to aid in research ideation and experiment design, to assist with the analysis and interpretation of experimental results, and to draft and polish parts of the paper's text. We did not use generative AI to generate synthetic data, to prove mathematical claims, or to author the experimental codebase beyond routine assistance. All AI-assisted content, including reported numbers, statistical analyses, and claims, was manually reviewed and verified against the experiment logs and the canonical statistical report (`runs/stats_report.md`) before submission. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility Statement

- **Environment**: `pip install env-510k` 0.1.2 (Gymnasium `510K-v0`, PettingZoo AEC, 86 tests); pass action is masked when illegal.
- **Baselines & analysis**: `baselines/` with fixed seeds, crash-tolerant chunked checkpoints, and per-chunk eval history; canonical configuration in `experiments/SPEC.md`.
- **Compute**: one GPU; ≈500 env-steps/s for MaskablePPO (≈40 min per 1M-step run; the 4-mode × 5-seed matrix ≈13 GPU-hours).
- **Data**: per-run `*.eval.json`, `*.metrics.json`, and `*.history.jsonl`, plus bot reference lines, coupling / team-conditioning / inferability analyses, and the canonical statistical report (`runs/stats_report.md`); a replayable, machine-verified trajectory sample (4 modes × 200 games) carries per-decision belief labels.

References

Balla, M., Long, G. E. M., Jeurissen, D., Goodman, J., Gaina, R. D., Perez-Liebana, D. PyTAG: Challenges and Opportunities for Reinforcement Learning in Tabletop Games. IEEE Conference on Games (CoG) 2023. arXiv:2307.09905.

Bard, N., Foerster, J., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., Bowling, M. The Hanabi Challenge: A New Frontier for AI Research. Artificial Intelligence 280, 2020.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T. L., Seshia, S. A., Abbeel, P., Dragan, A. On the Utility of Learning about Humans for Human-AI Coordination. NeurIPS 2019.

Agapiou, C., Vezhnevets, A. S., Duéñez-Guzmán, E. A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kummerfeld, E., et al. Melting Pot 2.0. arXiv:2211.13746, 2022.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., et al. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. Autonomous Agents and Multi-Agent Systems 36(26), 2022.

Lanctot, M., Lockhart, E., Timbers, F., Davoodi, S., Lewis, M., Cheng, B., De Bono, V., Graepel, T., Bachrach, Y. OpenSpiel: A Framework for Reinforcement Learning in Games. arXiv:1908.09453, 2019.

Suphx: Mastering Mahjong with Deep Reinforcement Learning. arXiv:2003.13590, 2020.

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., et al. Human-level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. Science 378(6624):1067–1074, 2022.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N. Stable-Baselines3: Reliable Reinforcement Learning Implementations. Journal of Machine Learning Research 22(268):1–8, 2021.

Zha, D., Lai, K.-H., Cao, Y., Huang, S., Wei, R., Guo, J., Hu, X. RLCard: A Toolkit for Reinforcement Learning in Card Games. AAAI-20 Workshop on Reinforcement Learning in Games, 2020. arXiv:1910.04376.

Samvelyan, M., Rashid, T., de Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., Whiteson, S. The StarCraft Multi-Agent Challenge. AAMAS 2019.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. Proximal Policy Optimization Algorithms. arXiv:1707.06347, 2017.

Terry, J. K., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., Williams, N., Lokesh, Y., Ravi, P. PettingZoo: Gym for Multi-Agent Reinforcement Learning. NeurIPS 2021 (Datasets & Benchmarks).

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., et al. Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032, 2024.

Zha, D., Xie, J., Ma, W., Zhang, S., Lian, X., Hu, X., Liu, J. DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning. ICML 2021.

Anonymous. Information-Induced Gradient Contraction in Partially-Observable MARL. In submission, 2027.

A CAPABILITY PROFILE AND PAIRED SAME-DEAL REFERENCE

First-place rate vs. rule-bot opponents with rule/random references (player 0, 500 reference games). Mean $\pm$ 95% CI over seeds. In SINGLE the agent's team is the agent alone, so finish-based team win and first place coincide.

Table 7: First-place rate (canonical obs-v2, MLP, 5 seeds).

mode	first-place	rule ref	random ref
SINGLE	0.262 ± 0.016	0.242	0.080
STATIC	0.280 ± 0.000	0.232	0.104
DYNAMIC	0.206 ± 0.069	0.236	0.112
OBVIOUS	0.362 ± 0.022	0.280	0.168

On the *same* 100 canonical deals, the trained policy finishes first more often than the rule bot at seat 0 in every mode (SINGLE 0.262 vs. 0.180, STATIC 0.280 vs. 0.160, DYNAMIC 0.206 vs. 0.170, OBVIOUS 0.362 vs. 0.210), while finish-based team-win shows no consistent gap on the same deals. The earlier "no method exceeds the reference" reading was an artifact of comparing against reference lines computed on a *different* 500-game sample; same-deal pairing matters more than sample size when the deal set moves absolute rates by ~ 0.1. Because the rule-bot outcomes are shared across policy seeds, we report these as a descriptive paired same-deal reference rather than as independent-seed significance tests.

B 1V3 VS 2V2 ROBUSTNESS BY SEED

Table 8: DYNAMIC first-place rate by red-team size (canonical obs-v2, 100 games/seed).

bucket	first	finish team-win
2v2	0.200 ± 0.091	0.503 ± 0.064
1v3	0.221 ± 0.020	0.314 ± 0.079