WHEN BODIES COMMIT: COMMITMENT-GROUNDED POLICY LEARNING IN EMBODIED GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents can remain strategically undecided after their bodies have already made some outcomes physically unreachable. In competitive interaction, this mismatch creates a distinctive vulnerability: an opponent can observe physical commitment and exploit choices that can no longer be reversed. We formalize this phenomenon through continuation capability and develop commitment-grounded policy learning, which adaptively expands both players' high-level strategy sets while keeping the underlying physical skills fixed. Our analysis connects the resulting finite policy-space game to the original embodied interaction, quantifies the approximation induced by a finite responder set, and yields an oracle-based certificate of the remaining strategic gap. On a humanoid–quadruped penalty system, fixed side-blind response sets underestimate learned responders by up to 18% once commitment becomes readable. A controlled game calibrated from the robot measurements reproduces this failure and enables multi-seed evaluation. Across eight seeds, adaptive responder-set expansion consistently improves over one-sided learning, archive-only minimax, and self-play, reducing the worst-case held-out responder payoff by 9% relative to self-play with twice the training budget. These results identify physical commitment as a strategic variable in embodied learning: robustness depends not only on what an agent intends to do, but also on what its body can still do.

1 INTRODUCTION

Most learning-in-games formulations separate strategy from execution. Embodied interaction violates that separation: momentum, contact, latency, and closed-loop whole-body control can erase a terminal alternative before a high-level policy stops assigning it probability, so the effective game depends on what the body can still execute, not only on what the policy can still command.

The humanoid–quadruped penalty system motivating this work makes the mismatch measurable: a 10 Hz game policy commands frozen soccer whole-body controllers (S-WBCs), shot-side reversibility is lost about 0.29 s before contact while the side is readable earlier, and changing only ball speed shifts the responder's value of deferring. Commitment thus changes the strategic problem, but that observation alone gives no learning algorithm with a guarantee.

We ask whether that structure can be turned into a strategy learner that carries a certificate, and then test the answer term by term. The design choice is to learn new *game-level policies*, not merely to reweight a frozen archive. Physical commitment supplies a small, executable responder library $\mathcal{B} = \{\beta_b\}_{b=1}^{B}$ such as left commitment, right commitment, and deferral. For any initiator game policy π, let

$$a_\theta(\pi)_b = J_\theta(\pi, \beta_b) \in [0, 1] \tag{1}$$

be the responder payoff under physical setting θ. At iteration k, a finite archive $\mathcal{A}_k = \{\pi_1, \ldots, \pi_{J_k}\}$ defines the restricted zero-sum game

$$V_k = \min_{w \in \Delta(J_k)} \max_{b \in [B]} (A_k w)_b, \qquad A_k = [a_\theta(\pi_1), \ldots, a_\theta(\pi_{J_k})]. \tag{2}$$

Its dual $q_k \in \Delta(B)$ is the mixture of executable responder behaviors currently hardest for the archive. We then train a *new* game policy against $b \sim q_k$ using the same terminal zero-sum reward as the original self-play system, freeze both S-WBCs, add the learned policy to the archive, and re-solve the game. Figure 1 shows the embodied system; Sec. 3 gives the loop.

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.

Whether solving and expanding this finite game means anything for the original interaction is settled by two results. Theorem 1 shows that for *every* initiator strategy ν, the finite-library objective $f_\theta(\nu)$ and the full-responder objective $F_\theta(\nu)$ satisfy $f_\theta(\nu) \le F_\theta(\nu) \le f_\theta(\nu) + \delta_{B,\theta}$, where the responder-library gap $\delta_{B,\theta}$ measures how well arbitrary responders are approximated by mixtures of the library; the abstraction error is explicit rather than assumed away. Theorem 2 then shows that the restricted value never increases when a column is added and that, if the oracle is an ϵ_k-best response to q_k and its new column has reduced cost $g_k = V_k - q_k^\top a_\theta(\pi_{k+1})$,

$$\boxed{0 \le V_k - V_B \le g_k + \epsilon_k, \qquad F_\theta(\nu_k) - V_\theta \le \delta_{B,\theta} + g_k + \epsilon_k,} \tag{3}$$

where V_B is the best value any mixture of game policies attains against the library and ν_k is the restricted primal mixture; Proposition 1 shows the oracle objective admits an exact likelihood-ratio gradient, so the only approximation in the practical actor update is the optimizer's, which is what ϵ_k measures.

Our contributions are three. **First**, we turn physical commitment into a falsifiable response abstraction and bound its effect on the full dynamic game for any finite executable library. **Second**, we derive commitment-grounded policy-space learning with an initiator oracle and a responder oracle, and prove monotonic restricted improvement and a two-sided exploitability certificate whose reduced-cost terms are measured in the loop and whose oracle-error terms we estimate against held-out evaluators (Corollary 1); the optimizer is a double oracle, the contribution is what it certifies about the embodied game. **Third**, we evaluate every term of that certificate on a humanoid–quadruped system and on a controlled game built from its measured constants (Sec. 5): a fixed side-blind library understates the worst-response value once commitment is readable (0.18 on the robot, 0.27 in the controlled game), and growing the library with a responder oracle lowers the held-out worst-response value below self-play at twice the budget, archive-only minimax, and one-sided learning in every seed; how the library is seeded does not matter.

2 Related Work

Differential games and one-sided information. Zero-sum differential games (Isaacs, 1965; Başar & Olsder, 1999) and their reachability formulations (Mitchell et al., 2005; Tomlin et al., 2000) characterize what each player can still enforce from a state; games of incomplete information study how a player with a private type controls what the opponent learns (Aumann & Maschler, 1995; De Meyer, 1996; Cardaliaguet, 2007), recently extended to state constraints and to a football case study (Ghimire et al., 2024; 2026). Our object differs: no type is realized at the start; the terminal choice stays open until dynamics remove a continuation. We target a trainable procedure with an exploitability certificate rather than a value solver (App. N).

Learning in games and policy populations. Equilibrium computation by best-response dynamics goes back to fictitious play and multiplicative weights (Fudenberg & Levine, 1998; Freund & Schapire, 1999); double-oracle methods grow both strategy sets by best responses (McMahan et al., 2003), a column-generation scheme (Dantzig & Wolfe, 1960; Desrosiers & Lübbecke, 2005). Their deep counterparts—fictitious self-play, PSRO and its variants, and league training—replace exact best responses by learned policies (Heinrich et al., 2015; Lanctot et al., 2017; Balduzzi et al., 2019; Muller et al., 2020; McAleer et al., 2020; Vinyals et al., 2019), with exploitability estimated by learned best responses (Lockhart et al., 2019; Timbers et al., 2022) and populations ranked by evolutionary dynamics (Omidshafiei et al., 2019). Self-play has reached superhuman play in board and card games (Silver et al., 2018; Moravčík et al., 2017; Brown & Sandholm, 2019; Perolat et al., 2022). Our loop is a double oracle over game-level policies; what is new is what its value and reduced costs certify about the embodied game (Sec. 4), and the finding that the responder oracle, not the initial rows, carries the gain (Sec. 5.3).

Embodied competition. Competitive self-play in physical simulation yields emergent skills and autocurricula (Bansal et al., 2018; Baker et al., 2020; Liu et al., 2019; Lowe et al., 2017; Yu et al., 2022), and legged-soccer systems demonstrate agile whole-body skills learned by imitation and RL (Peng et al., 2018; Haarnoja et al., 2024; Tirumala et al., 2025; Ji et al., 2023; Rudin et al., 2022). These optimize game return without asking when the body has eliminated an alternative; we keep a 10 Hz game actor over frozen whole-body controllers (App. I) and add no skill.

Commitment. Commitment as a chosen strategic device is classical (Schelling, 1960; Fudenberg & Tirole, 1985; Caruana & Einav, 2008), and learners that anticipate others' adaptation appear in strategic classification, performative prediction, and opponent shaping (Hardt et al., 2016; Perdomo et al., 2020; Foerster et al., 2018; Lu et al., 2022); we study commitment imposed by physics.

3 Commitment-Grounded Policy-Space Learning

3.1 Embodied game and physical commitment

Consider a finite-horizon two-player zero-sum dynamic game under physical parameters θ between an initiator S and a responder G, with the architecture of Fig. 1. Strategic commands are sampled from game-level policies,

$$c_t^S \sim \pi_{\phi_S}^{\text{game}}(\cdot \mid h_t^S), \tag{4}$$

$$c_t^G \sim \pi_{\phi_G}^{\text{game}}(\cdot \mid h_t^G, \hat{y}_t), \qquad \hat{y}_t = f_\omega(h_t^G), \tag{5}$$

and frozen S-WBCs map strategic commands to low-level actions

$$u_t^i = \pi_{\psi_i}^{\text{S-WBC}}(o_t^i, c_t^i), \qquad s_{t+1} \sim P_\theta(\cdot \mid s_t, u_t^S, u_t^G). \tag{6}$$

Only the game policies are learned in the proposed method; ψ_S, ψ_G and the intent estimator are frozen exactly as in the underlying self-play system. The responder payoff $\sigma(\tau) \in [0,1]$ is scored from the realized terminal physical event, with initiator payoff $-\sigma$.

Let $Y = g(s_{0:T}) \in \mathcal{Y}$ be a realized terminal outcome. For a state s_t, a declared class Π^{cont} of initiator continuation controllers, and a fixed responder probe π_G^{probe}, define continuation capability

$$C_t(y; \Pi^{\text{cont}}) = \sup_{\pi^{\text{cont}} \in \Pi^{\text{cont}}} \mathbb{P}(Y = y \mid s_t, \pi^{\text{cont}}, \pi_G^{\text{probe}}, \theta), \tag{7}$$

and the η-reachable set $\mathcal{R}_t^\eta = \{y : C_t(y) \geq \eta\}$.

Definition 1 (Persistent physical commitment)**.** For alternatives $A \subseteq \mathcal{Y}$, the commitment time is

$$t^c(A) = \inf\{t < T : A \not\subseteq \mathcal{R}_u^\eta \text{ for all } u \in [t, T]\}, \tag{8}$$

with $t^c = T$ if no persistent loss occurs.

This is a measurement relative to $(\eta, \Pi^{\text{cont}}, \pi_G^{\text{probe}})$, not a certified backward-reachable set. Figure 2 shows the robot instantiation and the deadlines that bound the no-reaction margin Δ.

3.2 Response library and full game

Let Π_S denote the admissible game-level initiator policy class under the frozen S-WBC, and Π_G the responder policy class. A commitment-grounded response library is a finite set

$$\mathcal{B} = \{\beta_1, \ldots, \beta_B\} \subseteq \Pi_G \tag{9}$$

of executable responder behaviors. In the penalty instantiation the intended minimal basis is left commitment, right commitment, and deferral; they must be implemented by forcing responder behavior, issued no earlier than the initiator's commitment time t^c, rather than by conditioning on what a learned responder happened to do; Sec. 5.2 shows what goes wrong when this timing condition is violated.

For any initiator policy $\pi \in \Pi_S$, define the response-payoff vector

$$a_\theta(\pi) = [J_\theta(\pi, \beta_1), \ldots, J_\theta(\pi, \beta_B)]^\top, \quad J_\theta(\pi, \pi_G) = \mathbb{E}[\sigma(\tau)]. \tag{10}$$

A distribution $\nu \in \Delta(\Pi_S)$ over initiator policies has finite-library and full-response objectives

$$f_\theta(\nu) = \max_{b \in [B]} \mathbb{E}_{\pi \sim \nu} a_\theta(\pi)_b, \tag{11}$$

$$F_\theta(\nu) = \sup_{\pi_G \in \Pi_G} \mathbb{E}_{\pi \sim \nu} J_\theta(\pi, \pi_G). \tag{12}$$

The corresponding values are

$$V_{B,\theta} = \inf_{\nu \in \Delta(\Pi_S)} f_\theta(\nu), \qquad V_\theta = \inf_{\nu \in \Delta(\Pi_S)} F_\theta(\nu). \tag{13}$$

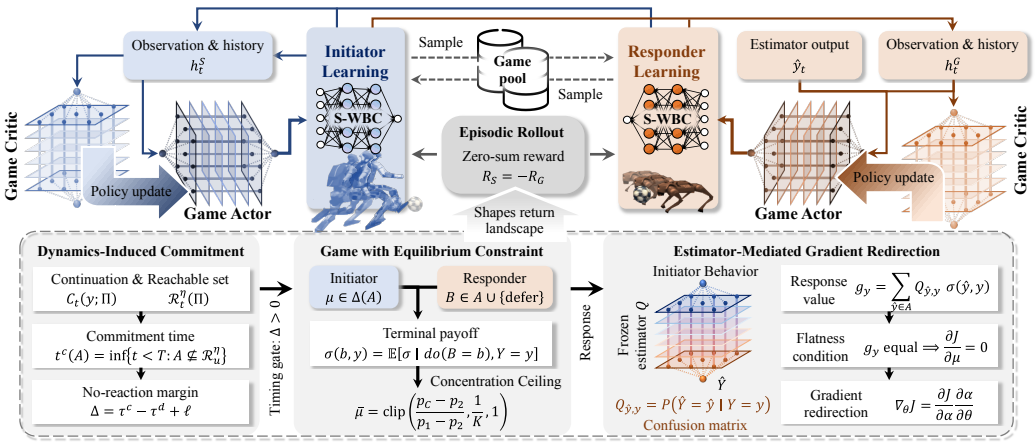


Figure 1: The embodied game. Top: initiator and responder game policies over frozen whole-body controllers (S-WBC) play under the terminal zero-sum reward and draw opponents from a shared pool. Bottom: the structure physical rollouts induce — commitment time t^c and margin Δ, a deferral-capped terminal game, and estimator-mediated flatness (App. F).

The response library is useful only insofar as it spans the strategic effect of more general responders. We quantify that instead of assuming it.

Definition 2 (Responder-library gap)**.**

$$\delta_{B,\theta} = \sup_{\pi_G \in \Pi_G} \inf_{q \in \Delta(B)} \sup_{\pi \in \Pi_S} \left| J_\theta(\pi, \pi_G) - q^\top a_\theta(\pi) \right|. \tag{14}$$

When δ_B is small, every responder acts on every initiator policy like some mixture of the library. We drop θ when the setting is fixed.

3.3 Restricted Game, Oracles, and Algorithm

At iteration k, let

$$\mathcal{A}_k = \{\pi_1, \ldots, \pi_{J_k}\} \subset \Pi_S, \qquad A_k = [a_\theta(\pi_1), \ldots, a_\theta(\pi_{J_k})] \in [0,1]^{B \times J_k}. \tag{15}$$

The restricted game is

$$V_k = \min_{w \in \Delta(J_k)} \max_b (A_k w)_b = \max_{q \in \Delta(B)} \min_{j \le J_k} q^\top A_{k,:,j}. \tag{16}$$

Let (w_k, q_k) be a saddle pair. The dual mixture q_k is not merely an evaluation object: it defines the policy-learning oracle objective

$$\mathcal{L}_k(\pi) = q_k^\top a_\theta(\pi) = \mathbb{E}_{b \sim q_k} J_\theta(\pi, \beta_b). \tag{17}$$

Because the initiator minimizes responder payoff, the exact oracle value is

$$m_k = \inf_{\pi \in \Pi_S} \mathcal{L}_k(\pi). \tag{18}$$

A practical game-policy learner returns π_{k+1} with oracle error

$$\epsilon_k = \mathcal{L}_k(\pi_{k+1}) - m_k \ge 0, \tag{19}$$

and reduced cost

$$g_k = V_k - \mathcal{L}_k(\pi_{k+1}). \tag{20}$$

A positive g_k means the learned policy exposes a direction unavailable to the current archive against the restricted-game dual. Symmetrically, a responder oracle trained against the deployed mixture ν_k returns β_{k+1} with reduced cost $h_k = J_\theta(\nu_k, \beta_{k+1}) - V_k$; Algorithm 1 alternates the two oracles and adds a column or row whenever its reduced cost exceeds a tolerance τ_{stop}, until neither does. Section 4 shows what the resulting reduced costs certify.

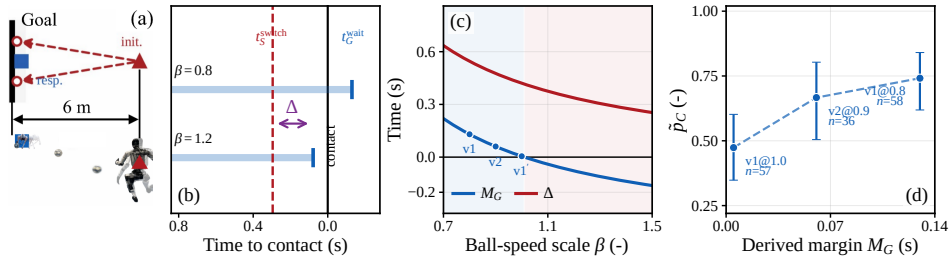


Figure 2: Robot instantiation and timing. (a) Humanoid initiator, quadruped responder, two low aim points. (b) Initiator's last switch time t_S^{switch} and responder's latest commit time t_G^{wait} at two ball speeds; Δ is where the responder must already have acted. (c) Coverage margin M_G and Δ against ball-speed scale β (derived from measured constants). (d) Measured deferral coverage against M_G, 95% Jeffreys intervals.

Algorithm 1 Commitment-grounded policy-space learning (one-sided: skip lines 6–7)

Require: initiator archive $\mathcal{A}_0$, commitment-grounded response library $\mathcal{B}_0$, frozen S-WBCs, setting θ, oracle budgets, tolerance τ_{stop}

1: **for** $k = 0, 1, \ldots$ **do**
2: Estimate A_k (rows $\mathcal{B}_k$, columns $\mathcal{A}_k$) with independent physical rollouts; solve Eq. 16 for (w_k, q_k, V_k).
3: Initiator oracle: from the prescribed checkpoint, sample $b \sim q_k$ per episode and optimize Eq. 17 under the original terminal payoff; call the result π_{k+1}.
4: On fresh episodes form $g_k = V_k - q_k^\top a_\theta(\pi_{k+1})$.
5: **if** $g_k > \tau_{\text{stop}}$ over its interval **then** $\mathcal{A}_{k+1} \leftarrow \mathcal{A}_k \cup \{\pi_{k+1}\}$.
6: Responder oracle: train a responder $\beta_{k+1} \in \Pi_G$ against the episode mixture ν_k; on fresh episodes form $h_k = J_\theta(\nu_k, \beta_{k+1}) - V_k$.
7: **if** $h_k > \tau_{\text{stop}}$ over its interval **then** $\mathcal{B}_{k+1} \leftarrow \mathcal{B}_k \cup \{\beta_{k+1}\}$.
8: **if** nothing was added **then** stop and return $(\mathcal{A}_k, w_k)$.
9: **end for**

4 Theoretical Analysis

4.1 Response Reduction

We first quantify what is lost when the full responder class is replaced by the library.

Theorem 1 (Uniform response reduction)**.** *For every $\nu \in \Delta(\Pi_S)$,*

$$\boxed{f_\theta(\nu) \le F_\theta(\nu) \le f_\theta(\nu) + \delta_{B,\theta}.} \tag{21}$$

Consequently,

$$0 \le V_\theta - V_{B,\theta} \le \delta_{B,\theta}. \tag{22}$$

If an initiator strategy ν is ε-optimal for the finite response game, $f_\theta(\nu) - V_{B,\theta} \le \varepsilon$, then its full dynamic-game exploitability satisfies

$$\boxed{F_\theta(\nu) - V_\theta \le \delta_{B,\theta} + \varepsilon.} \tag{23}$$

One abstraction error thus controls every initiator distribution (proof in App. A).

The global δ_B is not observable, but for any set of held-out responders the residual $d_g = \min_{q \in \Delta(B)} \|a_g - A^\top q\|_\infty$ is (Sec. 5.2), and App. B gives a sufficient decomposition $\delta_{B,\theta} \le \epsilon_{\text{term}} + \epsilon_{\text{mix}}$.

The symmetric concentration ceiling of the prior analysis of this system is the special case of symmetric alternatives (App. F); the learner requires neither symmetry nor a factorization through terminal-side frequencies.

4.2 The Oracle as a Policy Gradient

Proposition 1 (The oracle has an exact policy-gradient objective)**.** *Fix q_k and parameterize the initiator game policy by ϕ. If responder behavior $b \sim q_k$ is sampled once per episode, the S-WBC*

and physical dynamics do not depend on ϕ except through the game-policy actions, and the responder payoff is the terminal random variable $\sigma(\tau)$, then

$$\nabla_\phi \mathcal{L}_k(\phi) = \mathbb{E}_{b\sim q_k,\, \tau\sim\pi_\phi,\beta_b}\left[\sigma(\tau)\sum_t \nabla_\phi \log \pi_\phi(c_t^S \mid h_t^S)\right]. \tag{24}$$

Subtracting any history-dependent baseline that is independent of the sampled strategic action preserves the expectation.

Thus the existing game-level actor–critic serves as the oracle without changing the S-WBC; PPO clipping makes it approximate, and its residual is the ϵ_k term below.

4.3 Monotone improvement and the two-sided certificate

The theorem is stated for exact payoffs; its finite-sample version (Corollary 2) is in App. E. Let

$$\mathcal{C}_\theta = \mathrm{cl}\,\mathrm{conv}\{a_\theta(\pi) : \pi \in \Pi_S\} \subset [0,1]^B, \tag{25}$$

which is compact by construction after closure. The full finite-library value can then be written as

$$V_{B,\theta} = \min_{x\in\mathcal{C}_\theta} \max_b x_b = \max_{q\in\Delta(B)} \inf_{\pi\in\Pi_S} q^\top a_\theta(\pi). \tag{26}$$

Theorem 2 (Monotonic policy-space improvement and oracle-gap certificate)**.** *At iteration k, let (w_k, q_k, V_k) solve the restricted game of Eq. 16, let π_{k+1} satisfy Eq. 19, and let g_k be Eq. 20. Then:*

(i) Monotonicity. *After adding π_{k+1},*

$$\boxed{V_{k+1} \le V_k.} \tag{27}$$

(ii) Finite-library optimality gap.

$$\boxed{0 \le V_k - V_{B,\theta} \le g_k + \epsilon_k.} \tag{28}$$

With an exact oracle, $g_k = 0$ certifies the archive mixture optimal against $\mathcal{B}$.

(iii) Embodied dynamic-game lift. *Let ν_k sample $\pi_j \in \mathcal{A}_k$ according to w_k. Then*

$$\boxed{F_\theta(\nu_k) - V_\theta \le \delta_{B,\theta} + g_k + \epsilon_k.} \tag{29}$$

Abstraction, restricted-game gap, and oracle error enter additively; with a finite Π_S and an exact oracle the loop terminates after at most $|\Pi_S|$ additions.

Two-sided certificate. Theorem 1 never requires the library to be fixed or hand-designed: any finite $\mathcal{B}_k \subseteq \Pi_G$ of executable behaviors is admissible. Growing $\mathcal{B}_k$ with a responder oracle (lines 6–7) replaces the global, unobservable δ_B by a reduced cost measured for the deployed mixture.

Corollary 1 (Two-sided certificate)**.** *Let $\mathcal{B}_k \subseteq \Pi_G$ be the current library, ν_k the restricted primal mixture, and let the responder oracle return β_{k+1} with error $\epsilon'_k = F_\theta(\nu_k) - J_\theta(\nu_k, \beta_{k+1}) \ge 0$. With g_k and ϵ_k as in Theorem 2 for the library $\mathcal{B}_k$,*

$$\boxed{F_\theta(\nu_k) - V_\theta \;\le\; g_k + h_k + \epsilon_k + \epsilon'_k.} \tag{30}$$

The proof (App. D) replaces $\delta_{B,\theta}$ in Eq. 29 by the exploitability of the current mixture relative to the grown library, which the responder oracle measures as h_k. Both reduced costs are observable; only the oracle errors remain.

5 Results and Discussion

5.1 Physical evidence for commitment

Figure 5 shows the interaction and Fig. 3 the four robot measurements the formulation rests on.

Under the declared continuation-controller class and $\eta = 0.5$, persistent loss of shot-side reversibility occurs at 0.642 m of foot–ball separation, about 0.29 s before contact at the measured swing speed, while a linear probe on body state already reads the final shot side at about 0.92 accuracy 0.5 s

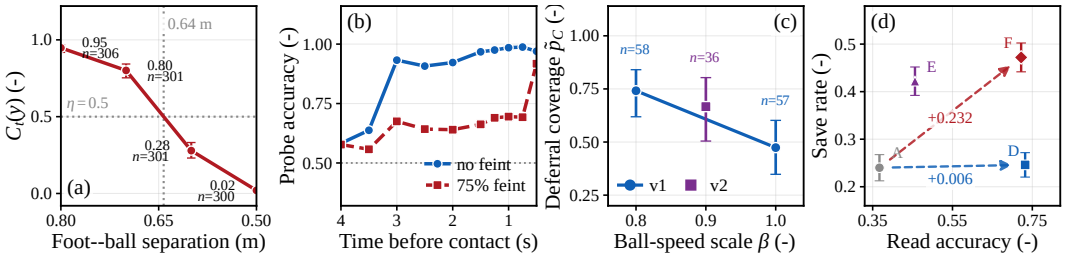


Figure 3: Robot measurements, 95% Jeffreys intervals. (a) Continuation capability against foot–ball separation; the $\eta = 0.5$ crossing is 0.29 s before contact. (b) Linear-probe read accuracy with and without a 75% feint rate. (c) Deferral coverage under a ball-speed-only intervention. (d) Read accuracy against save rate for four responder variants; arrows are one-factor moves.

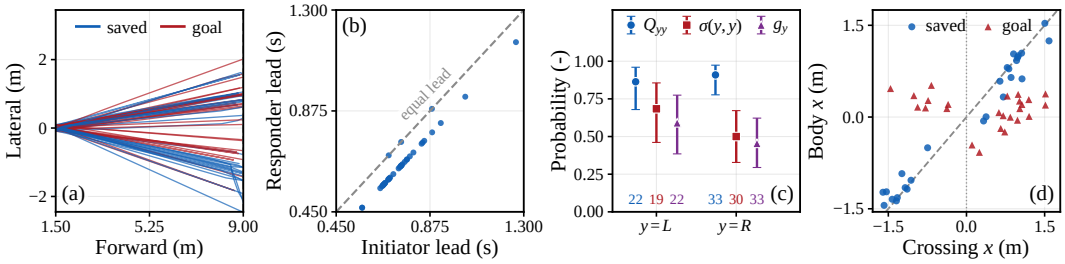


Figure 4: State recording, policy F. (a) Ball tracks by terminal event. (b) Responder lead against initiator lead: every episode is below the diagonal, so the responder commits after the initiator. (c) Read accuracy Q_{yy}, matched payoff $\sigma(y, y)$ and response value g_y (difference unresolved, Fisher $p = 0.41$). (d) Ball crossing against final responder position.

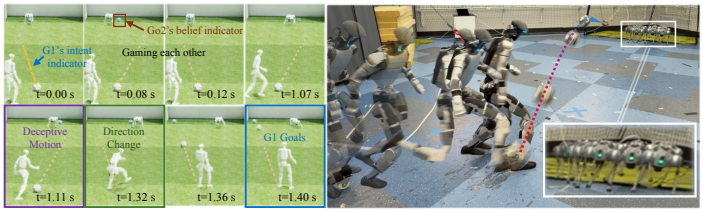


Figure 5: A feint in simulation (left) and on hardware (right): the humanoid switches side just before contact and scores. Indicators are for visualisation; an illustration, not evidence.

out. Commitment therefore closes after the outcome becomes predictable, which is what creates a readable-but-still-reversible interval to reason about; the state recording confirms that the responder's read locks only after the ball's side is fixed (Fig. 4b). Changing only ball speed, with both learned policies fixed, moves the responder's deferral coverage from 0.741 ($n = 58$) at $\beta = 0.8$ to 0.474 ($n = 57$) at $\beta = 1.0$: the value of waiting is a property of the physics, not only of the policies. Across responder variants (924–1081 episodes each), nearly equal gains in read accuracy ($0.37 \to 0.73$ and $0.37 \to 0.72$) buy $+0.006$ and $+0.232$ in save rate (Fig. 3d). Prediction accuracy is thus not a proxy for game value, so the library is defined by executable behavior, not estimator quality.

5.2 One-sided learning on the robot

We first ran the one-sided loop (Algorithm 1 without lines 6–7) on the robot (App. J): a five-snapshot archive from one self-play league at $\beta = 0.9$, forced L/R/defer rows, a 7.37M-transition budget per condition, and held-out evaluation on disjoint seed pools with environment-cluster bootstrap intervals (one algorithm seed per condition; App. L).

Minimax over the archive lowers the held-out worst case to 0.248 $[0.216, 0.282]$, from 0.376 for the initial snapshot and 0.420 for matched-budget self-play (Fig. 6b). The oracle's column, however, received zero weight, and every trained column had negative held-out reduced cost (Fig. 6c), so the certificate correctly refused to credit it. The raw episodes show why (Fig. 6d): the forced keeper is

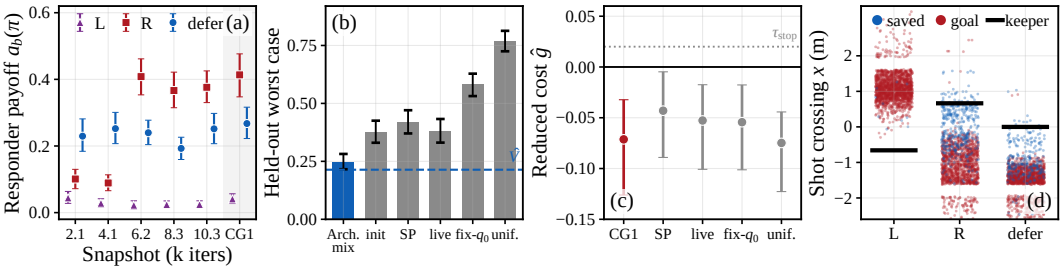


Figure 6: Robot audit, E1–E2 at $\beta = 0.9$, held-out episodes (lower is better for the initiator). (a) Payoff under forced L, R, defer for five archive snapshots and the loop's column CG1. (b) Worst case of the archive mixture, the initial snapshot, and four matched-budget fine-tunes; dashed: $\hat{V}$. (c) Held-out reduced cost of every trained column; $\tau_{\text{stop}} = 0.02$. (d) Shot crossing under each forced response, with the forced keeper's position at contact.

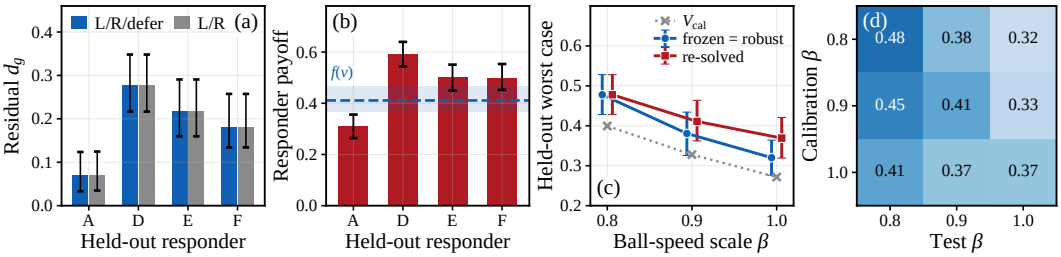


Figure 7: Robot audit, E3–E4 on a three-policy library. (a) Coverage residual of four held-out learned responders, full library vs L/R only. (b) Learned-responder payoff against the library's minimax strategy, with the forced worst case $f(\nu)$. (c) Held-out worst case across ball speeds, frozen, re-solved, or robust weights. (d) Calibration $\times$ test worst-case matrix.

Table 1: Controlled commitment game, round 4 (eight seeds). Held-out worst-response value $\hat{F}$ (responder payoff against the strongest of six held-out evaluators; lower is better), the reduction achieved by two-sided learning, and the number of seeds in which it is lower. Mean with 95% t-interval over seeds.

Method	$\hat{F} \downarrow$	Reduction by ours	Seeds
Two-sided policy-space learning (ours)	**0.850** [0.838, 0.862]	—	—
Same loop seeded by a learned responder (PSRO)	0.851 [0.836, 0.865]	0.001 [−0.013, 0.015]	5/8
One-sided policy-space learning	0.890 [0.882, 0.897]	0.040 [0.022, 0.058]	8/8
Archive-only minimax	0.899 [0.889, 0.909]	0.049 [0.038, 0.060]	8/8
Self-play, matched budget	0.911 [0.900, 0.923]	0.061 [0.042, 0.080]	8/8
Self-play, 2× budget	0.940 [0.934, 0.945]	0.090 [0.077, 0.102]	8/8

already stationary at ∓ 0.66 m at contact, and the shot crosses on the far side in 99.6% and 87.9% of forced-L and forced-R episodes, so the executed rows were visible before the initiator's 0.29 s commitment. Learned responders exceed the library's forced worst case by up to 0.18 (Fig. 7b; App. K): a side-blind library cannot represent responders that *read* the initiator's commitment.

5.3 Controlled commitment game

To isolate that failure mode we built a minimal game in the spirit of Ghimire et al. (2024) with every constant taken from the robot (App. G): full aim commits 0.29 s before contact, the responder reads the aim with a 0.10 s delay at 0.97 accuracy, and post-contact timing noise reproduces the two measured deferral coverages (0.740, 0.473 vs. 0.741, 0.474). Both sides are linear-softmax policies trained by REINFORCE on the terminal payoff; the library is L and R issued *at* t^c, defer, and follow-the-read. Experiments were run in four rounds with hypotheses and seeds fixed before each round, and all rounds are reported (App. G); intervals are 95% t-intervals over algorithm seeds.

In rounds 1–2 the one-sided loop converges in one to three columns with $\hat{g}_k$ reaching zero, at a forced worst case of 0.609 [0.607, 0.610], 0.012 above the defer floor that no initiator can beat. It beats

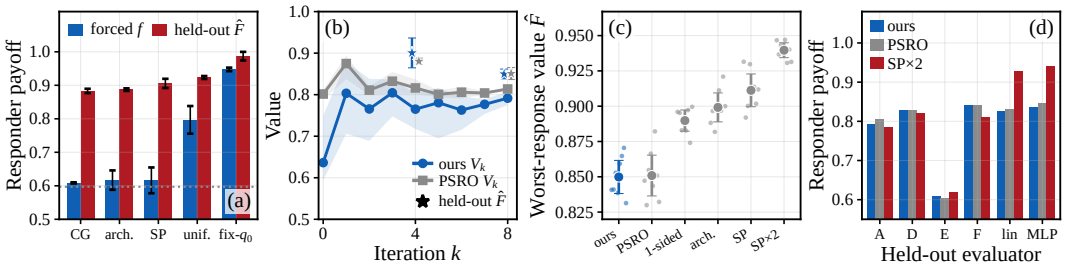


Figure 8: Controlled commitment game. (a) Rounds 1–2, one-sided methods: forced worst case f against held-out $\hat{F}$; dotted: defer floor. (b) Round 4: restricted value V_k for ours and PSRO (mean, seed range) with $\hat{F}$ after four iterations and at termination (stars). (c) Round 4 $\hat{F}$ by method, seeds as dots, 95% t-intervals. (d) Payoff of each held-out evaluator against the final strategy.

uniform-response and fixed-q_0 training (0.797, 0.948; $5/5$ seeds) but not archive-only minimax or self-play, already near the floor. Held-out learned responders, however, reach 0.883 against it: the library understates the worst-response value by 0.275 $[0.269, 0.281]$, and by 0.25–0.29 for every method, library variant, and commitment timing tested (Fig. 8a). This is the robot's failure, isolated: once commitment is readable before contact, fixed side-blind rows cannot span responders that act on the read.

Growing the library with a responder oracle (Corollary 1) lowered $\hat{F}$ below self-play, one-sided learning, and archive-only minimax in $5/5$ seeds in round 3 (App. G). Round 4 repeated it on eight new seeds against self-play with *twice* the budget and the same loop seeded with a learned responder (PSRO), evaluated by a 32-unit MLP responder from a class no method trains against, the linear best response, and four scripted responders. Table 1 reports the outcome: the two-sided strategy is the least exploitable method, below self-play at twice the budget by 0.090 and below the other baselines by 0.040–0.061, in $8/8$ seeds each, with every interval excluding zero (Fig. 8c). Longer self-play is *worse* ($0.911 \to 0.940$): chasing the current responder makes it more readable (Fig. 8d). The loop adds four to eight responder rows, and its restricted value rises from 0.64 toward 0.8 as the library learns to read (Fig. 8b). At termination $\hat{g}_k + \hat{h}_k$ undershoots $\hat{F} - V$ by 0.006–0.026: the oracle errors of Corollary 1 are not negligible, since the linear responder oracle is weaker than the held-out MLP.

Seeded with the league's learned responder instead of commitment-timed rows, the loop reaches the same worst-response value (Table 1; -0.020 $[-0.054, 0.014]$ after four iterations). The gain comes from learning the responder set; physical commitment enters through the theory, which explains why a fixed side-blind library must fail and what the certificate must measure. Re-solving across ball speeds beats a frozen $\beta = 0.8$ solution by only 0.042 $[-0.012, 0.096]$, and the dual-mixture oracle's advantage over uniform and live training has a five-seed interval including zero; we claim neither (App. M lists every hypothesis with its outcome).

6 Conclusion

Embodied games are constrained not only by what a policy chooses, but also by what the underlying physical system can still execute. This work formalizes that distinction through physical commitment and develops commitment-grounded policy learning around the evolution of executable alternatives. By combining continuation capability with an adaptive policy-space game, the proposed framework expands both players' high-level strategy sets while preserving the underlying physical skills, and its theoretical analysis makes the approximation induced by a finite responder set explicit while providing an oracle-based certificate of the remaining strategic gap. Experiments on the humanoid–quadruped system show that fixed side-blind response sets become inadequate once commitment is readable, and a controlled game calibrated from the robot measurements reproduces the same failure while enabling systematic multi-seed evaluation. Across these settings, adaptively expanding the responder set improves robustness over one-sided learning, archive-only minimax, and self-play. More broadly, the results suggest that strategic reasoning in embodied intelligence should be grounded in the evolution of executable possibilities, rather than policy outputs alone, positioning physical commitment as a bridge between game-theoretic learning and the irreversible structure imposed by real dynamics.

AI USE STATEMENT

Generative AI tools were used for manuscript editing and presentation and to assist parts of the consistency checking.

REPRODUCIBILITY STATEMENT

The code and data will be open-sourced immediately upon acceptance. The assumptions and complete proofs for the response reduction and oracle certificates are given in Apps. A–E. The controlled-game dynamics, training budgets, evaluators, and seed-level certificate diagnostics are specified in App. G; the robotic hierarchy, measurements, executed evaluation protocol, and statistical procedures are specified in Apps. I, **??**, and O.

REFERENCES

Robert J. Aumann and Michael Maschler. Repeated games with incomplete information. *MIT Press*, 1995.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020.

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 434–443, 2019.

Trapit Bansal, Jakub Pachocki, Szymon Sidor, Ilya Sutskever, and Igor Mordatch. Emergent complexity via multi-agent competition. In *International Conference on Learning Representations*, 2018.

Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 2nd edition, 1999.

Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Pierre Cardaliaguet. Differential games with asymmetric information. *SIAM Journal on Control and Optimization*, 46(3):816–838, 2007.

Guillermo Caruana and Liran Einav. A theory of endogenous commitment. *The Review of Economic Studies*, 75(1):99–116, 2008.

George B. Dantzig and Philip Wolfe. Decomposition principle for linear programs. *Operations Research*, 8(1):101–111, 1960.

Bernard De Meyer. Repeated games, duality and the central limit theorem. *Mathematics of Operations Research*, 21(1):51–69, 1996.

Jacques Desrosiers and Marco E. Lübbecke. A primer in column generation. In *Column Generation*, pp. 1–32. Springer, 2005.

Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2018.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1–2):79–103, 1999.

Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

Drew Fudenberg and Jean Tirole. Preemption and rent equalization in the adoption of new technology. *The Review of Economic Studies*, 52(3):383–401, 1985.

Mukesh Ghimire, Lei Zhang, Zhe Xu, and Yi Ren. State-constrained zero-sum differential games with one-sided information. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15512–15539. PMLR, 2024. URL https://proceedings.mlr.press/v235/ghimire24a.html.

Mukesh Ghimire, Lei Zhang, Zhe Xu, and Yi Ren. Solving football by exploiting equilibrium structure of 2p0s differential games with one-sided information. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0df745d55f0c45ffc3d408a27cb1785f-Paper-Conference.pdf.

Tuomas Haarnoja, Ben Moran, Guy Lever, Sandy H. Huang, Dhruva Tirumala, Jan Humplik, Markus Wulfmeier, Saran Tunyasuvunakool, Noah Y. Siegel, Roland Hafner, Michael Bloesch, Kristian Hartikainen, Arunkumar Byravan, Leonard Hasenclever, Yuval Tassa, Fereshteh Sadeghi, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 9(89), 2024. doi: 10.1126/scirobotics.adi8022. URL https://arxiv.org/abs/2304.13653.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Innovations in Theoretical Computer Science (ITCS)*, 2016.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 805–813. PMLR, 2015. URL https://proceedings.mlr.press/v37/heinrich15.html.

Rufus Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. Wiley, 1965.

Yandong Ji, Gabriel B. Margolis, and Pulkit Agrawal. DribbleBot: Dynamic legged manipulation in the wild. In *IEEE International Conference on Robotics and Automation*, pp. 5155–5162, 2023.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3323fe11e9595c09af38fe67567a9394-Abstract.html.

Siqi Liu, Guy Lever, Josh Merel, Saran Tunyasuvunakool, Nicolas Heess, and Thore Graepel. Emergent coordination through competition. In *International Conference on Learning Representations*, 2019. URL https://research.google/pubs/emergent-coordination-through-competition/.

Edward Lockhart, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, and Karl Tuyls. Computing approximate equilibria in sequential adversarial games by exploitability descent. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 464–470, 2019.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Christopher Lu, Timon Willi, Christian A. Schroeder de Witt, and Jakob N. Foerster. Model-free opponent shaping. In *International Conference on Machine Learning (ICML)*, 2022.

Stephen McAleer, John Lanier, Roy Fox, and Pierre Baldi. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 536–543, 2003.

Ian M. Mitchell, Alexandre M. Bayen, and Claire J. Tomlin. A time-dependent hamilton–jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005. doi: 10.1109/TAC.2005.851439.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2020.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9:9937, 2019.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4):143:1–143:14, 2018.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, 2020.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning (CoRL)*, pp. 91–100, 2022.

Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pp. 3487–3493, 2022.

Dhruva Tirumala, Markus Wulfmeier, Ben Moran, Sandy Huang, Jan Humplik, Guy Lever, Tuomas Haarnoja, Leonard Hasenclever, Arunkumar Byravan, Nathan Batchelor, Neil Sreendra, Kushal Patel, Marlon Gwira, Francesco Nori, Martin Riedmiller, and Nicolas Heess. Learning robot soccer from egocentric vision with deep reinforcement learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 165–184. PMLR, 2025. URL https://proceedings.mlr.press/v270/tirumala25a.html.

Claire J. Tomlin, John Lygeros, and S. Shankar Sastry. A game theoretic approach to controller design for hybrid systems. *Proceedings of the IEEE*, 88(7):949–970, 2000.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019. doi: 10.1038/s41586-019-1724-z.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

A Proof of Theorem 1

Every proof in this appendix fixes a single physical setting θ and suppresses it from the notation; θ reappears only in App. I, where the dependence on physics is itself the object of study. Fix θ and abbreviate $J(\pi, g) = J_\theta(\pi, g)$, $f = f_\theta$, and $F = F_\theta$. Because every forced response β_b is itself an admissible responder policy, $\mathcal{B} \subseteq \Pi_G$. Hence for every initiator distribution ν,

$$F(\nu) = \sup_{g \in \Pi_G} \mathbb{E}_{\pi\sim\nu} J(\pi, g) \geq \max_b \mathbb{E}_{\pi\sim\nu} J(\pi, \beta_b) = f(\nu). \tag{31}$$

For the reverse inequality, fix any $g \in \Pi_G$ and any $\varepsilon > 0$. By Definition 2, there exists $q_g \in \Delta(B)$ such that

$$\sup_{\pi \in \Pi_S} \left| J(\pi, g) - q_g^\top a(\pi) \right| \leq \delta_B + \varepsilon. \tag{32}$$

Taking expectation over $\pi \sim \nu$ gives

$$\mathbb{E}_{\pi\sim\nu} J(\pi, g) \leq q_g^\top \mathbb{E}_{\pi\sim\nu} a(\pi) + \delta_B + \varepsilon \tag{33}$$

$$\leq \max_b \mathbb{E}_{\pi\sim\nu} a_b(\pi) + \delta_B + \varepsilon = f(\nu) + \delta_B + \varepsilon. \tag{34}$$

Taking the supremum over g and then $\varepsilon \downarrow 0$ yields Eq. 21. Minimizing both sides gives

$$V_B \leq V \leq V_B + \delta_B, \tag{35}$$

which is Eq. 22. If $f(\nu) \leq V_B + \varepsilon$, then

$$F(\nu) - V \leq f(\nu) + \delta_B - V_B \leq \delta_B + \varepsilon, \tag{36}$$

proving Eq. 23. □

B Terminalization Bound for the Responder-Library Gap

The global δ_B is not directly observable, but the commitment interpretation yields a falsifiable sufficient condition. Suppose a responder trajectory can be assigned to one of the declared terminal behavior classes. Against initiator policy π, let $q_\pi(\pi_G) \in \Delta(B)$ be the resulting behavior distribution. Define

$$\epsilon_{\text{term}} = \sup_{\pi_G, \pi} \left| J_\theta(\pi, \pi_G) - q_\pi(\pi_G)^\top a_\theta(\pi) \right|, \tag{37}$$

$$\epsilon_{\text{mix}} = \sup_{\pi_G} \inf_{q \in \Delta(B)} \sup_{\pi \in \Pi_S} \|q_\pi(\pi_G) - q\|_1. \tag{38}$$

Then

$$\delta_{B,\theta} \leq \epsilon_{\text{term}} + \epsilon_{\text{mix}}. \tag{39}$$

The first term tests whether terminal behavior labels retain the payoff effect of a general responder; the second tests whether a single episode-level mixture explains that responder consistently across initiator policies. We report these quantities only when the logged trajectory makes the terminal behavior label unambiguous.

Proof. Suppose the terminalization map is well-defined. Fix a responder g and initiator policy π. By definition of ϵ_{term},

$$J(\pi, g) \leq q_\pi(g)^\top a(\pi) + \epsilon_{\text{term}}. \tag{40}$$

For any common mixture $q \in \Delta(B)$,

$$J(\pi, g) - q^\top a(\pi) \leq [q_\pi(g) - q]^\top a(\pi) + \epsilon_{\text{term}} \tag{41}$$

$$\leq \|q_\pi(g) - q\|_1 \|a(\pi)\|_\infty + \epsilon_{\text{term}} \tag{42}$$

$$\leq \|q_\pi(g) - q\|_1 + \epsilon_{\text{term}}, \tag{43}$$

using $a_b(\pi) \in [0, 1]$. Applying the same argument to the negative difference yields

$$\left| J(\pi, g) - q^\top a(\pi) \right| \leq \|q_\pi(g) - q\|_1 + \epsilon_{\text{term}}. \tag{44}$$

Taking $\sup_\pi$, then $\inf_q$, then $\sup_g$ gives Eq. 39. The decomposition is conservative: terminal behavior may be payoff-sufficient even when the raw mixture distributions vary more strongly than required.

C Proof of Proposition 1

For fixed q_k, sample $b \sim q_k$ once at episode start and let $p_\phi(\tau \mid b)$ denote the trajectory density under the game policy and forced response behavior. The oracle objective is

$$\mathcal{L}_k(\phi) = \sum_b q_{k,b} \int \sigma(\tau) p_\phi(\tau \mid b)\, d\tau. \tag{45}$$

The physical transition kernel and frozen S-WBC do not depend on ϕ except through sampled strategic actions, hence

$$\nabla_\phi \log p_\phi(\tau \mid b) = \sum_t \nabla_\phi \log \pi_\phi(c_t^S \mid h_t^S). \tag{46}$$

Applying the likelihood-ratio identity gives Eq. 24. For a baseline $b_t(h_t^S)$ independent of c_t^S conditional on history,

$$\mathbb{E}\left[b_t(h_t^S) \nabla_\phi \log \pi_\phi(c_t^S \mid h_t^S)\right] = 0, \tag{47}$$

so standard advantage baselines preserve the expectation. PPO's clipping changes the exact objective and is therefore treated as an approximate optimizer, not as part of the identity. □

C.1 Finite-library minimax identity

Let $\mathcal{C} = \mathrm{cl}\,\mathrm{conv}\{a(\pi) : \pi \in \Pi_S\} \subset [0,1]^B$. Since $\mathcal{C}$ and $\Delta(B)$ are compact convex sets and $q^\top x$ is bilinear,

$$V_B = \min_{x \in \mathcal{C}} \max_{q \in \Delta(B)} q^\top x = \max_{q \in \Delta(B)} \min_{x \in \mathcal{C}} q^\top x. \tag{48}$$

A linear functional takes the same infimum over $\mathcal{C}$ as over its generating policy set, hence

$$V_B = \max_{q \in \Delta(B)} \inf_{\pi \in \Pi_S} q^\top a(\pi). \tag{49}$$

C.2 Monotonicity

The archive convex hull at iteration k is contained in the archive convex hull at $k+1$. Minimizing the same convex objective $x \mapsto \max_b x_b$ over a superset cannot increase the optimum, so $V_{k+1} \le V_k$.

C.3 Oracle-gap certificate

Let q_k be an optimal dual mixture of the restricted game and

$$m_k = \inf_{\pi \in \Pi_S} q_k^\top a(\pi). \tag{50}$$

Equation 49 implies $m_k \le V_B$, because q_k is one feasible responder mixture in the maximization defining V_B. Restricted optimization can only be worse for the initiator than optimization over the full policy class, so $V_B \le V_k$. Therefore

$$0 \le V_k - V_B \le V_k - m_k. \tag{51}$$

The ϵ_k-best-response condition gives

$$q_k^\top a(\pi_{k+1}) \le m_k + \epsilon_k, \tag{52}$$

so $m_k \ge q_k^\top a(\pi_{k+1}) - \epsilon_k$. Substitution into Eq. 51 yields

$$V_k - V_B \le V_k - q_k^\top a(\pi_{k+1}) + \epsilon_k = g_k + \epsilon_k. \tag{53}$$

This proves part (ii).

For part (iii), let ν_k be the episode-level mixture induced by w_k. Its finite-library payoff is exactly $f(\nu_k) = V_k$. By Theorem 1,

$$F(\nu_k) \le V_k + \delta_B. \tag{54}$$

Moreover $V \ge V_B$ because the full responder class contains the forced response library. Hence

$$F(\nu_k) - V \le V_k + \delta_B - V_B \tag{55}$$

$$\le \delta_B + g_k + \epsilon_k. \tag{56}$$

If the oracle is exact and $g_k = 0$, the inequality above gives $V_k = V_B$. For a finite policy class, every exact-oracle iteration with positive unresolved gap can add a policy not already sufficient for the restricted optimum; after at most $|\Pi_S|$ distinct additions, no unseen policy remains and the exact finite-library equilibrium is reached. □

D Proof of Corollary 1

Write $f_k(\nu) = \max_{b \in \mathcal{B}_k} \mathbb{E}_{\pi\sim\nu} J(\pi, \beta_b)$ and $V_{\mathcal{B}_k} = \inf_\nu f_k(\nu)$. Because $\mathcal{B}_k \subseteq \Pi_G$, $V_{\mathcal{B}_k} \le V$ (the argument of App. A). The restricted mixture satisfies $f_k(\nu_k) = V_k$, and Theorem 2(ii) applied to the library $\mathcal{B}_k$ gives $V_k - V_{\mathcal{B}_k} \le g_k + \epsilon_k$. By definition of the responder oracle error, $F(\nu_k) = J(\nu_k, \beta_{k+1}) + \epsilon'_k = V_k + h_k + \epsilon'_k$. Hence

$$F(\nu_k) - V \le F(\nu_k) - V_{\mathcal{B}_k} = h_k + \epsilon'_k + (V_k - V_{\mathcal{B}_k}) \le g_k + h_k + \epsilon_k + \epsilon'_k. \tag{57}$$

□

E Finite-Sample Version of the Oracle Certificate

Corollary 2 (Finite-sample oracle certificate)**.** *Let $\hat{A}_k$ be an estimated restricted matrix with $\|\hat{A}_k - A_k\|_\infty \le \eta_A$, let $(\hat{w}_k, \hat{q}_k, \hat{V}_k)$ be an exact saddle point of the estimated game, and let the newly trained policy have an independently estimated payoff vector $\hat{a}_{k+1}$ with $\|\hat{a}_{k+1} - a_\theta(\pi_{k+1})\|_\infty \le \eta_{\text{new}}$. Suppose π_{k+1} is an ϵ_k-best response to the* estimated *dual mixture $\hat{q}_k$, and define*

$$\hat{g}_k = \hat{V}_k - \hat{q}_k^\top \hat{a}_{k+1}. \tag{58}$$

Then, on the joint confidence event,

$$\boxed{F_\theta(\hat{\nu}_k) - V_\theta \le \delta_{B,\theta} + [\hat{g}_k]_+ + \eta_A + \eta_{\text{new}} + \epsilon_k,} \tag{59}$$

where $\hat{\nu}_k$ samples the current archive using $\hat{w}_k$. Solver feasibility residuals can be added directly to the right-hand side.

The oracle is trained against $\hat{q}_k$, the dual that appears in the bound, so no assumption is made that the estimated dual equals the unknown exact one. The proof (App. **??**) is a column-generation argument: q_k makes the best-response value m_k a lower bound on $V_{B,\theta}$, so the reduced cost of the newly *learned* column measures how much optimality may still be missing, up to ϵ_k.

What the learning step must add. If line 3 is removed, V_k cannot improve after the initial restricted solve; newly trained columns are the only way to expand the feasible payoff hull. Whether they do is empirical, and the held-out reduced cost is the test: a column with $\hat{g}_k \le \tau_{\text{stop}}$ does not enter. E1 therefore compares the full loop with archive-only minimax under one protocol and reports the held-out reduced cost of every new column.

Why train against the dual. On the estimator-following branch of the prior analysis, live-responder return can be first-order flat to terminal reallocation (App. F); the dual mixture instead targets the response the archive currently handles worst, and E2 compares it with uniform and live training from identical checkpoints.

We prove Corollary 2 and specify the intervals used by the experimental protocol.

Let $\hat{A}$ satisfy $\|\hat{A} - A\|_\infty \le \eta_A$. For any mixture w,

$$\left| \max_b (\hat{A}w)_b - \max_b (Aw)_b \right| \le \eta_A, \tag{60}$$

because every row payoff changes by at most $\eta_A \sum_j w_j = \eta_A$. Applying this at the estimated optimum $\hat{w}_k$, where $\max_b (\hat{A}\hat{w}_k)_b = \hat{V}$ by construction, gives the quantity the certificate needs — the *true* finite-library payoff of the mixture the algorithm actually returns:

$$f(\hat{\nu}_k) = \max_b (A\hat{w}_k)_b \le \hat{V} + \eta_A. \tag{61}$$

(The same inequality applied at w_k gives $V_k \le \hat{V} + \eta_A$, which is not needed below.) Now fix the estimated dual $\hat{q}$ used to train the oracle. The full finite-library value satisfies

$$V_B \ge \inf_{\pi \in \Pi_S} \hat{q}^\top a(\pi). \tag{62}$$

If the learned policy is an ϵ_k-best response to $\hat{q}$, then

$$\inf_{\pi} \hat{q}^\top a(\pi) \ge \hat{q}^\top a(\pi_{k+1}) - \epsilon_k. \tag{63}$$

If $\|\hat{a}_{k+1} - a(\pi_{k+1})\|_\infty \le \eta_{\text{new}}$, then

$$\hat{q}^\top a(\pi_{k+1}) \ge \hat{q}^\top \hat{a}_{k+1} - \eta_{\text{new}}. \tag{64}$$

Combining with Eq. 61,

$$f(\hat{\nu}_k) - V_B \le \hat{V} + \eta_A - [\hat{q}^\top \hat{a}_{k+1} - \eta_{\text{new}} - \epsilon_k] \tag{65}$$

$$= \hat{g}_k + \eta_A + \eta_{\text{new}} + \epsilon_k. \tag{66}$$

Replacing $\hat{g}_k$ by $[\hat{g}_k]_+$ only weakens the bound. Theorem 1 adds δ_B, proving Eq. 59.

Simultaneous bounded-payoff intervals. If every cell contains $n_{b,j}$ independent payoffs in $[0, 1]$, Hoeffding plus a union bound over BJ cells gives

$$\eta_A = \max_{b,j} \sqrt{\frac{\log(2BJ/\delta)}{2n_{b,j}}} \tag{67}$$

with probability at least $1 - \delta$. Binary save/no-save cells may instead use Bonferroni-corrected Clopper–Pearson intervals. The implementation records cellwise radii rather than collapsing them unless a scalar theorem display is needed.

Oracle uncertainty. The theorem intentionally does not pretend that finite rollouts certify the global neural best-response error ϵ_k. The experiment audits it by (i) at least three independent oracle-training seeds from the same checkpoint, (ii) a longer-budget oracle on selected iterations, and (iii) direct comparison against uniform-response and live-responder training. The best observed held-out oracle value is a diagnostic upper bound on m_k, not a formal lower bound. Final wording must preserve this distinction.

F Structural Theory of the Robotic System as Special Cases

F.1 Symmetric concentration ceiling

For K symmetric terminal alternatives, a matched response gives p_1, a mismatched response p_2, and deferral p_C, with $p_2 \le p_C \le p_1$ and $p_2 < p_1$. For terminal allocation μ let $m(\mu) = \max_y \mu(y)$ and $\delta = p_1 - p_2$. The responder best-response value is

$$W(\mu) = \max\{p_C, p_2 + \delta m(\mu)\}. \tag{68}$$

Since $m(\mu) \ge 1/K$, the game value is

$$V^\star = \max\{p_C, p_2 + \delta/K\}. \tag{69}$$

Writing $r = (p_C - p_2)/\delta$, optimality is equivalent to

$$m(\mu) \le \bar{\mu} = \text{clip}(r, 1/K, 1), \tag{70}$$

and

$$\text{Expl}(\mu) = \delta[m(\mu) - \bar{\mu}]_+. \tag{71}$$

This result explains the original equilibrium geometry but is not required by Algorithm 1.

F.2 Estimator-mediated flatness

Let the responder follow an outcome estimator with confusion matrix $Q_{\hat{y},y}$ and terminal response payoff $\sigma(\hat{y}, y)$. Define

$$g_y(Q) = \sum_{\hat{y}} Q_{\hat{y},y} \sigma(\hat{y}, y), \qquad J(\mu, Q) = -\sum_y \mu(y) g_y(Q). \tag{72}$$

The direct derivative along the simplex tangent space vanishes if and only if g_y is constant across outcomes. Under the symmetric model and a symmetric estimator of accuracy α, $g_y \equiv \alpha p_1 + (1-\alpha)p_2$ and the estimator-following threshold is $\alpha > r$. When $\alpha \geq 1/K$, the effective threshold equals the concentration ceiling, $\alpha^\star_{\text{eff}} = \bar{\mu}$. This is a mechanism result, not a convergence theorem. In the present paper it motivates E2: a dual-mixture oracle should provide a clearer strategic training objective precisely where live-responder return may offer weak direct reallocation pressure.

Figure 9 illustrates both results and the tournament evidence that first motivated them. In a scripted-feint world (75% switch rate, switch at 0.8 m) deferring to the ball scores 0.736 against 0.502 for following the early read of the body; in the reactive world of the common protocol the estimator-replacement policy F scores highest. The two worlds use different opponents and policy sets, so the comparison establishes an ordering of responder policies, not exploitability.

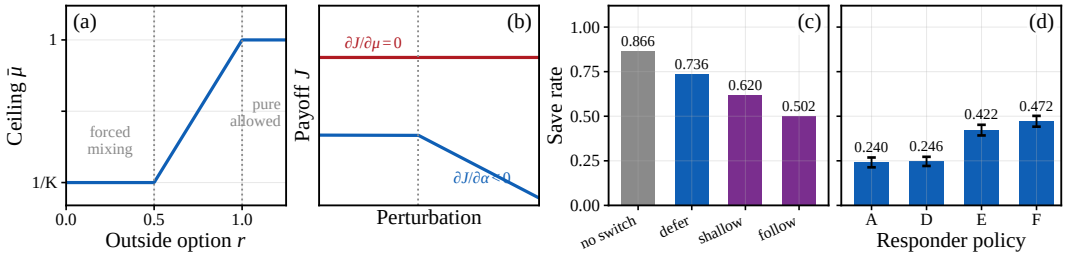


Figure 9: Structural theory of the robotic system and the responder tournaments. (a) Schematic, no data: the outside option $r = (p_C - p_2)/\delta$ caps admissible terminal concentration at $\bar{\mu} = \text{clip}(r, 1/K, 1)$; below $r = 1/K$ the initiator is forced to mix, above $r = 1$ a pure aim is allowed. (b) Schematic, no data: through an outcome estimator the payoff is flat along reallocations μ when response values agree, while estimator accuracy α beyond r moves it. (c) Save rate of four responder policies in the scripted-feint world; the no-switch world is a reference. (d) Save rate of the four responder policies of Fig. 13 in the reactive world (924–1081 episodes each), with 95% Jeffreys intervals.

F.3 Why the factorization $A = SM$ is optional

For forced behavior b, initiator policy j, and terminal outcome Y,

$$A_{b,j} = \sum_y \mathbb{P}(Y = y \mid \pi_j, \beta_b)\mathbb{E}[\sigma \mid \pi_j, \beta_b, Y = y]. \tag{73}$$

The simpler factorization $A = SM$ requires both terminal payoff sufficiency given (b, Y) and invariance of the initiator's terminal-outcome distribution to the forced responder behavior. Neither condition is assumed by the direct policy-space learner. E3 tests them as optional structure.

G Controlled Commitment Game

Dynamics. Time runs over the last 1.0 s before contact in 0.05 s decision steps. The initiator's aim $a \in [-0.85, 0.85]$ m moves by $\{-1, 0, +1\} \times \omega\, dt$ per step with $\omega = 0.85/0.19$ m s^{-1} and is frozen in the last 0.10 s, so a full aim is irreversible 0.29 s before contact (the measured t^c). The shot crosses at $a_T + \mathcal{N}(0, 0.06^2)$. The responder moves by $\{-1, 0, +1\} \times 1.6\, dt$ per step within ± 0.85 m, sees the aim with a 0.10 s delay and $\mathcal{N}(0, 0.45^2)$ noise (read accuracy 0.97 for a full aim), and after contact chases the ball; it saves iff $t_f - 0.12 - \max(|x - k| - 0.25, 0)/1.6 + \xi \geq 0$ with $t_f = 0.5/\beta$ and $\xi \sim \mathcal{N}(-0.0164, 0.176^2)$, the only calibrated quantity, fitted to the two measured deferral coverages. The initiator observes its aim, the responder's position, and time. Both sides use linear-softmax game policies over ten (initiator) and nine (responder) features, trained by REINFORCE with a batch-mean baseline and Adam (learning rate 0.05) on the terminal payoff alone.

Protocol. Experiments were run in four rounds; the hypotheses, methods, seeds, and decision rules of each round were fixed before it ran, and every round is reported. Rounds 1–2 (seeds 0–4) study one-sided learning. Each seed trains a self-play league for 600 iterations of 2048 episodes; its final snapshot is π_0. Oracle calls use 200 iterations of 2048 episodes, baselines the maximum

loop budget, and all evaluations use disjoint random streams with $20{,}000$ episodes per cell. Held-out worst-response value $\hat{F}$ is the maximum over four scripted responders (reactive follower, late follower, bait-and-defer, sharp follower) and a best-response responder trained from scratch (two restarts of 300 iterations of 4096 episodes). Reduced-cost intervals hold the dual fixed across bootstrap replicates, as Corollary 2 requires. Round 3 (two-sided learning, seeds 10–14) uses held-out $\hat{F}$ as its single primary metric; the loop runs at most eight iterations with a responder oracle of 200 iterations of 4096 episodes.

Table 2: Rounds 1–2 (round 2 shown; five seeds, $\beta = 0.9$, library L/R/defer/follow at t^c unless stated).

Method	forced worst case f	held-out worst-response value $\hat{F}$
CG, start $\{\pi_0\}$	0.609	0.883
CG, start five snapshots	0.611	0.881
archive-only $\{\pi_0\}$	0.617	0.887
archive-only, five snapshots	0.609	0.887
continued self-play	0.616	0.906
uniform-response PG	0.797	0.923
fixed-q_0 PG	0.948	0.987
CG, library L/R/defer	0.625	0.879
CG, commitments at 1.0 s	0.621	0.887

Table 3: Round 3 (seeds 10–14): held-out worst-response value $\hat{F}$ with 95% t-intervals.

Method	forced worst case f	held-out worst-response value $\hat{F}$
two-sided (ours)	0.648	0.855 $[0.843, 0.867]$
one-sided	0.607	0.878 $[0.866, 0.890]$
archive-only	0.621	0.886 $[0.876, 0.896]$
self-play	0.634	0.909 $[0.892, 0.926]$

Round 4. Eight seeds (20–27) compare the round-3 loop, the same loop seeded with the league's learned responder (PSRO), matched and doubled self-play, one-sided learning, and archive-only minimax, with anytime evaluation after four iterations. The evaluator adds a 32-unit MLP best-response responder over twelve features (aim read at two lags, read velocity, position, time, and interactions), two restarts of 400 iterations of 4096 episodes chosen on validation, and a linear best response with four restarts, to the four scripted responders. The MLP or the scripted sharp follower attains $\hat{F}$ for every method.

Certificate check. Corollary 1 bounds $F(\nu) - V$ by $g_k + h_k + \varepsilon_k + \varepsilon'_k$. Table 4 compares the terms the loop measures, $\hat{g}_k + \hat{h}_k$ at termination, with the gap the held-out evaluators measure, $\hat{F} - V$. For the two-sided loop the gap exceeds $\hat{g}_k + \hat{h}_k$ in the seven seeds that stopped before the iteration cap, by 0.006–0.026 (seed 20 hit the cap with $\hat{g}_k + \hat{h}_k = 0.226$); for PSRO it exceeds it in seven of eight seeds, by 0.019–0.041. The excess is an empirical lower estimate of the oracle errors $\varepsilon_k + \varepsilon'_k$, dominated by the responder oracle: the in-loop linear best response is weaker than the held-out MLP and sharp-follower evaluators that attain $\hat{F}$. The certificate is therefore sound only with these terms retained; a practitioner should not read $\hat{g}_k + \hat{h}_k$ alone as an exploitability bound. This diagnostic was added after the runs.

Table 4: Round 4: in-loop certificate versus held-out gap, per seed. V: restricted value at termination; $\hat{g} + \hat{h}$: last measured reduced costs; $\hat{F} - V$: held-out worst-response value minus V.

	two-sided (ours)			PSRO		
seed	V	$\hat{g} + \hat{h}$	$\hat{F} - V$	V	$\hat{g} + \hat{h}$	$\hat{F} - V$
20	0.775	0.226	0.066	0.814	0.050	0.034
21	0.788	0.041	0.053	0.795	0.005	0.035
22	0.787	0.059	0.066	0.803	0.027	0.051
23	0.777	0.035	0.061	0.814	0.028	0.053
24	0.806	0.047	0.053	0.800	0.015	0.044
25	0.798	0.050	0.066	0.809	0.054	0.073
26	0.787	0.057	0.083	0.800	0.031	0.051
27	0.779	0.036	0.052	0.790	0.001	0.042

Hypothesis outcomes. In rounds 1–2, one-sided learning does not beat archive-only minimax or self-play (0.008 $[-0.021, 0.037]$ and 0.008 $[-0.031, 0.046]$), because the initial policy is already 0.02 above the defer floor; the dual-mixture oracle beats uniform and live training by 0.29 $[-0.06, 0.64]$ and 0.28 $[-0.04, 0.60]$, the two ties being seeds whose q_0 is the defer row; the library underestimates exploitability by $\hat{F} - f = 0.275$; and re-solving after a dynamics shift gains 0.042 $[-0.012, 0.096]$ at $\beta = 1.0$. Round 3's primary hypothesis — $\hat{F}$ of two-sided learning below self-play, one-sided learning, and archive-only minimax, each interval excluding zero with at least $4/5$ seeds — holds with $5/5$ seeds for all three.

H Qualitative Rollouts in Simulation and on Hardware

Figure 5 (Sec. 5) shows a feint that scores; Figs. 10 and 11 show the responder reading the final side and saving. All three are selected illustrations, not evidence for any claim in the paper.

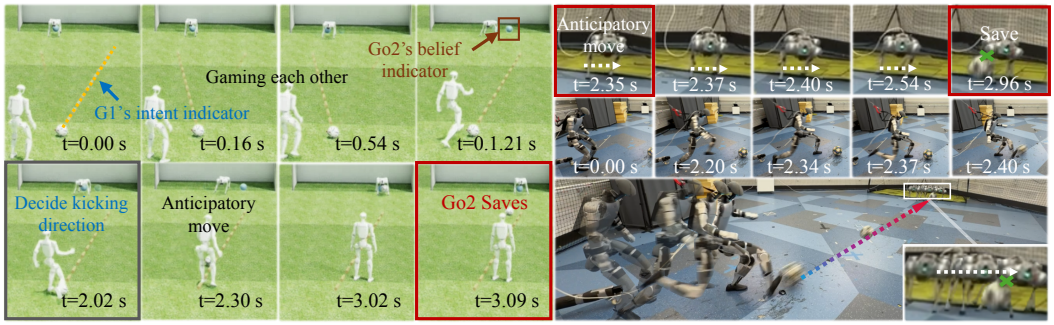


Figure 10: Anticipatory goalkeeping during continuous strategic interaction: the quadruped reads the humanoid's final shot side despite a late revision and saves. Left: simulation. Right: separately selected, vertically unsynchronised hardware key frames for each robot (top two rows) and the hardware overview (bottom). Intent and belief indicators are drawn for visualisation only.



Figure 11: Hardware deployment of anticipatory goalkeeping: the quadruped predicts the final shot side during kick preparation, starts its interception before contact, and blocks the ball.

I Robotic System and Existing Measurement Ledger

I.1 Hierarchy

The robotic system separates strategic policies from fixed soccer whole-body controllers. The humanoid shooter is initialized from self-collected human motion-capture data, retargeted to robot-feasible references, trained by motion imitation, and adapted by reinforcement learning. The quadruped saving controller is learned from scratch by reinforcement learning. During strategic self-play the low-level controllers and responder estimator are frozen while game-level actor–critics update against a policy snapshot pool. The proposed learner operates only on the game layer and does not average neural-network parameters or retrain the S-WBCs.

Figure 2 shows the instantiation and the timing model that the response library is built on, and Figs. 13 and 4 the two audits of the responder policies that the paper's measurements come from.

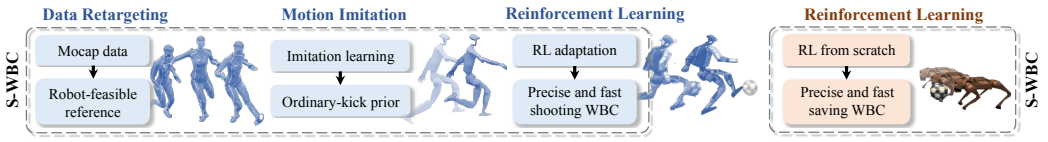


Figure 12: Training of the soccer whole-body controllers (S-WBC), frozen throughout this paper. Humanoid shooter: motion-capture data are retargeted to robot-feasible references, motion imitation yields an ordinary-kick prior, and RL adaptation turns it into a fast, precise shooting controller. Quadruped goalkeeper: RL from scratch yields a fast, precise saving controller. The game-level policies of Fig. 1 command these controllers at 10 Hz.

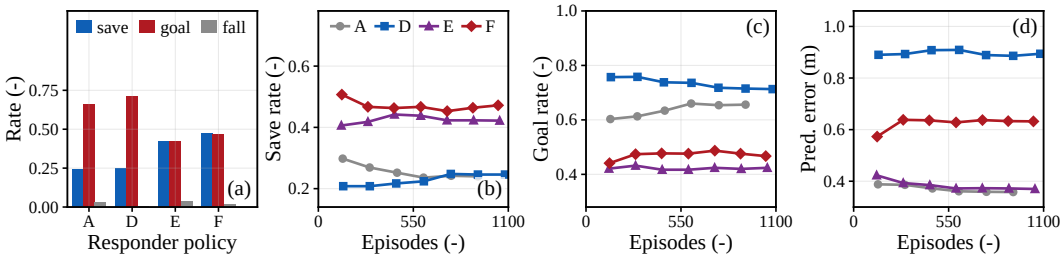


Figure 13: The four responder policies under one common protocol (924–1081 episodes each). (a) The full terminal-event distribution per policy, not only its save rate. (b–d) Cumulative save rate, goal rate, and kick-time prediction error against episode count, showing that each channel is stationary over the evaluation window. Policies D and F differ from the baseline A in a single factor each, which is what makes the comparison in Fig. 3(d) one-factor.

I.2 Measurement ledger

Table 5 is the complete list of robot measurements the paper relies on.

Table 5: Load-bearing measurements of the robotic system.

Quantity	Value	Status / interpretation
Continuation capability at 0.8/0.7/0.6/0.5 m	0.948/0.801/0.279/0.020	finite continuation-class measurement
Commitment boundary at $\eta = 0.5$	0.642 m; 0.29 s lead	measured lead upper-bounds full admissible lead
Readability at 0.5 s	about 0.92	prediction precedes measured commitment
v1 deferral coverage, $\beta = 0.8$	0.741 ($n = 58$)	fixed-policy physical setting
v1 deferral coverage, $\beta = 1.0$	0.474 ($n = 57$)	same v1 policy pair, changed ball speed
Policy A save/read	0.240 / 0.367 ($n = 924$)	common reactive-initiator protocol
Policy D save/read	0.246 / 0.732 ($n = 1081$)	deferral variant
Policy E save/read	0.422 / 0.456 ($n = 1050$)	deferral + decoy
Policy F save/read	0.472 / 0.722 ($n = 1041$)	estimator replacement
Policy F class-conditional diagonal accuracy	0.864 / 0.909	recorded confusion information
Policy F response values	$g_L = 0.591$, $g_R = 0.455$	difference unresolved, Fisher $p = 0.41$
Observational $\tilde{p}_2$	0.667 ($n = 3$)	too sparse for strong causal equilibrium claim

I.3 Observational equilibrium comparison

The two analyzed league runs have dominant terminal concentrations 0.850 (v1) and 0.917 (v2). Observational coverage proxies are $\tilde{p}_1 = 0.762$ ($n = 21$), $\tilde{p}_C^{v1} = 0.741$ ($n = 58$), $\tilde{p}_C^{v2} = 0.667$ ($n = 36$), and $\tilde{p}_2 = 0.667$ ($n = 3$). Propagating the prior posterior uncertainty (Fig. 14) gives $\Pr(m > \bar{\mu}) = 0.49$ for v1 and 0.78 for v2; equivalently, v1 sits inside its optimal set with probability 0.51, while v2 exceeds its ceiling with probability 0.78. The posterior probability that the v1 ceiling is at least the v2 ceiling is 0.87. These are directional observations, not identified interventional equilibrium tests, and they are not used as the direct training matrix of the new learner.

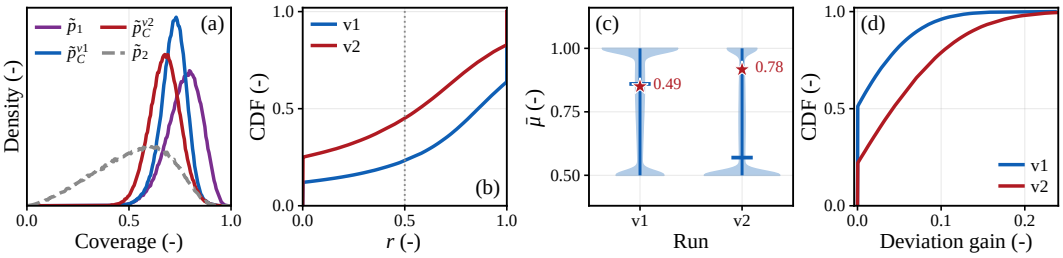


Figure 14: Posterior propagation for the two observational league runs, from Jeffreys Beta posteriors on the measured coverage cells (p_1 and p_2 shared across runs, p_C per run, draws with $p_2 \geq p_1$ rejected). (a) Posteriors of the four payoff cells; the sparse $\tilde{p}_2$ cell ($n = 3$) is what limits identification. (b) Posterior CDF of the outside-option ratio $r = (p_C - p_2)/(p_1 - p_2)$. (c) Posterior of the concentration ceiling $\bar{\mu}$ per run (violin and 95% interval, median as a bar) against the observed terminal concentration m (star); the printed number is $\Pr(m > \bar{\mu})$. (d) Posterior CDF of the predicted deviation gain. The v1 statement is not resolved at this sample size.

I.4 League training dynamics

The archive policies used by the robot audit come from two self-play league runs (v1 at $\beta = 0.8$, v2 at $\beta = 0.9$), one seed each (Fig. 15). Live save rate is not a stationary capability measure, because both opponents change; save rate against the frozen archive is, and it rises from 0.25 (v1) and 0.19 (v2) at the first evaluation to 0.67 and 0.66 over the last tenth of training. The initiator's switch count rises to about 6 per episode in v1 and 29 in v2 without any switching reward, so the feints the response library must handle emerge from the zero-sum terminal payoff alone. With one seed per run we make no claim about the distribution of training outcomes.

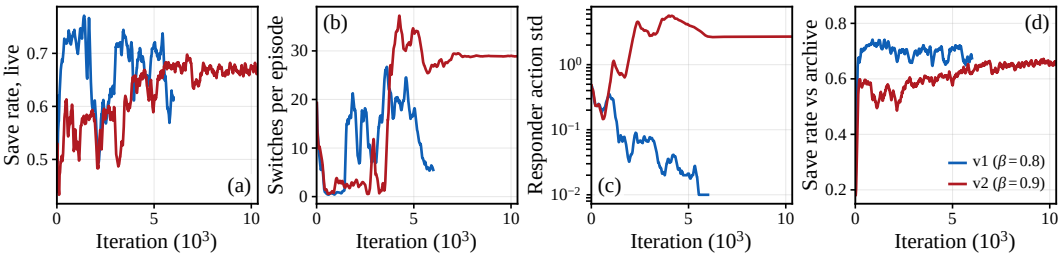


Figure 15: Self-play league training, one seed per run, 9-point moving averages. (a) Save rate, live responder against live initiator. (b) Initiator side switches per episode. (c) Responder action standard deviation, log scale. (d) Save rate against the frozen archive, the one capability measure not confounded by the opponent improving.

J Robot Evaluation Protocol

Table 6 summarizes the robot evaluation; App. L records the executed settings.

Table 6: Robot evaluation design. Each block targets one term of the certificate or the learning step itself.

ID	Question	Required comparison	Primary outputs
E1	Does learning new policies reduce worst-case game value?	proposed column generation vs archive-only minimax, ordinary live self-play, uniform-response PPO, and fixed-first-dual PPO; identical game-actor transition budget	V_k, held-out worst-case payoff, reduced cost g_k, new-policy success, support size
E2	Is the best-response oracle actually doing useful optimization?	from identical checkpoints: dual-mixture q_k vs uniform forced responses vs live responder; ≥ 3 seeds; short/long oracle budgets	$q_k^\top a(\pi)$, g_k, learning curves, gradient/update diagnostics
E3	Is the response abstraction adequate and which component matters?	L/R/defer vs L/R only; held-out learned responders A/D/E/F; direct payoff model vs symmetric/factorized model	coverage residuals, held-out worst-case payoff, failure cases
E4	Does the learner adapt when physical dynamics change?	train/re-solve at $\beta \in \{0.8, 0.9, 1.0\}$; frozen nominal, adaptive, and finite-set robust variants	cross-dynamics matrix, worst-case payoff, number of added policies, transferred reduced cost

J.1 E1: END-TO-END POLICY-SPACE LEARNING

Start from a fixed initial archive of chronologically spaced initiator snapshots and one designated trainable actor checkpoint. Implement forced left, right, and defer responder modes without modifying the S-WBC. Estimate the initial payoff matrix on training seeds, solve the restricted game, and train a new game actor against the dual mixture q_k. Use the same training-transition budget for all oracle baselines. Run at least three policy-expansion iterations or stop early if the held-out reduced-cost interval overlaps the predeclared tolerance. Primary plots: V_k vs actor-training transitions, held-out worst-case payoff, g_k with confidence interval, and support / archive growth.

J.2 E2: ORACLE ABLATION

For each iteration, clone the same starting actor into at least three random seeds for each oracle condition: proposed q_k, uniform forced-response mixture, ordinary live-responder self-play, and a fixed first-iteration dual mixture. The terminal reward and PPO implementation remain identical. Evaluate every resulting policy against all forced response rows with fresh seeds. Report q_k-weighted payoff, worst-row payoff, training return, gradient norm if already logged, and policy KL from the starting checkpoint. A longer-budget proposed-oracle run on at least one iteration audits whether short-budget optimization is the dominant source of ϵ_k.

J.3 E3: RESPONSE COVERAGE AND STRUCTURE

Use learned responders A/D/E/F strictly as held-out responders. Evaluate each against every policy added by the algorithm and compute the distance of its payoff vector to the convex hull of the L/R/defer basis. Repeat with the defer row removed. If trajectories admit unambiguous terminal behavior labels, also estimate sampled ϵ_{term} and ϵ_{mix}. Test $A = SM$ only as a structural ablation; the direct learner remains the primary method regardless of the outcome.

J.4 E4: DYNAMICS SHIFT

At $\beta = 0.8, 0.9, 1.0$, rebuild the forced-response payoff matrix using the same archive semantics. Compare the nominal 0.8 strategy frozen across settings, per-setting adaptive policy-space learning, and a robust variant whose restricted game stacks all measured (β, b) rows. The final plot is a train-setting $\times$ test-setting worst-case payoff matrix together with the number of new policies required after the shift.

K ROBOT AUDIT: RESPONSE COVERAGE AND DYNAMICS SHIFT

Figure 7 completes the robot audit of Sec. 5.2. Against a fixed three-policy library (the 6.2k snapshot and both CG columns), the held-out learned responders sit 0.07 (A), 0.28 (D), 0.22 (E), and 0.18 (F) from the convex hull of the forced payoff vectors, and three of them exceed the library's forced worst case of 0.411 $[0.365, 0.465]$ by up to 0.18. Across ball speeds the calibration value falls from 0.399 to 0.272, but re-solving the weights per setting is worse on held-out data (0.411 and 0.369 at $\beta = 0.9, 1.0$ against 0.381 and 0.320 frozen), because the three policies differ by at most 0.064 in any row.

L ROBOT EXPERIMENTS: EXECUTED SETTINGS

We list the settings under which the robot experiments of Sec. 5.2 were executed, including where they are narrower than the design of App. J.

Physical setting. E1–E3 were run at $\beta = 0.9$, the setting of the self-play league whose snapshots form the archive, rather than at $\beta = 0.8$.

Seeds. Every training condition — the column-generation loop, continued self-play, live-responder PPO, fixed-q_0 PPO, and uniform PPO — was run with one algorithm seed. No interval in the robot results reflects training-seed variability.

Loop length and budgets. Each oracle call ran 300 PPO updates from the 10.3k snapshot. The loop added one column and rejected the second at its stopping rule; the baselines ran 600 updates (7.37M actor transitions), matched to the loop's total. The long-budget oracle audit was not run.

Forced-response timing. The forced L/R behaviors move the responder to ∓ 0.66 m before the initiator's commitment point (Sec. 5.2). This timing makes the forced response observable to the initiator, which Sec. 5.2 identifies as the cause of the library's failure.

E3 and E4 library. E3 and E4 were evaluated on a fixed three-policy library (the 6.2k snapshot and the two columns the loop produced, including the one its stopping rule rejected), not on the final E1 archive, with 150 calibration and 300 test episodes per cell and per setting. No policy was retrained at $\beta = 0.8$ or 1.0. The E3 residuals and learned-responder payoffs are reported as computed by the evaluation pipeline; the supervised-responder episodes were not re-analysed independently.

M Hypotheses and Outcomes

Table 7 lists every hypothesis, the evidence required for it, and the outcome.

Table 7: Every hypothesis, the evidence required, and the outcome. Robot rows use the robot protocol (App. J); controlled-game rows use the protocols of App. G.

Claim	Required evidence	Outcome
Humanoid–quadruped robot		
Measured physical commitment precedes contact	robot measurement with controller-class and η caveats	supported (source)
Physical dynamics alter the game with policies fixed	same-policy change of deferral coverage / game value	supported: V_{cal} $0.399 \to 0.272$ over $\beta \in [0.8, 1.0]$
Minimax over the archive lowers worst-case payoff	held-out worst case below single policies	supported: 0.248 vs 0.376 (init), 0.420 (self-play)
Proposed learner beats archive-only minimax (G1)	repeatable held-out improvement over archive-only and self-play	**not supported**: deployed mixture identical to archive-only
Dual-mixture PPO is a useful oracle (G2)	lower q_k-weighted payoff / positive reduced cost vs alternatives	**not supported**: all $\hat{g} < 0$; uniform objective $0.217 \to 0.343$
Forced basis approximates learned responders (G3)	small held-out residuals relative to effects	**not supported**: d_g up to 0.28
Defer response is necessary	L/R/defer beats L/R on coverage or worst case	not supported: residuals unchanged
Re-solving across dynamics helps (G4)	adaptive/robust benefit under a measured shift	**not supported**: re-solved worse at $\beta = 0.9, 1.0$
Global dynamic-game near-optimality	credible δ_B and ϵ_k audits	not claimed
Controlled commitment game		
One-sided CG beats archive-only / self-play (sim G1)	5 seeds	**not supported**: no headroom above the defer floor
Fixed side-blind library spans learned responders (sim G3)	$\hat{F} - f$ small	**not supported**: 0.275
Two-sided learning lowers held-out worst-response value (round 3)	$\hat{F}$ below self-play, one-sided, archive-only; 5 seeds	**supported**: -0.054, -0.023, -0.031; $5/5$ seeds each
Two-sided beats $2\times$ self-play, strong evaluator (round 4, H1)	8 seeds, MLP evaluator	**supported**: -0.090 $[-0.102, -0.077]$, $8/8$
Two-sided beats one-sided, archive-only, self-play (round 4, H3)	8 seeds	**supported**: -0.040, -0.049, -0.061; $8/8$ each
Commitment-timed rows beat learned seeding (round 4, H2)	8 seeds	**not supported**: 0.001 $[-0.013, 0.015]$
Re-solving across dynamics helps (sim G4)	gain with interval excluding 0	not supported: 0.042 $[-0.012, 0.096]$

N Relation to State-Constrained One-Sided-Information Differential Games

Table 8 sets the two lines of work side by side.

O Statistical Reporting and Reproducibility

Units and splits. Each completed evaluation episode is one bounded-payoff trial. Seeds are assigned to training-matrix, oracle-training, oracle-held-out, and final-held-out pools before the corresponding

Table 8: The theories are complementary rather than nested. We do not claim to generalize the Hamilton–Jacobi results.

Axis	Ghimire et al. (Ghimire et al., 2024)	This paper
Strategic uncertainty	private payoff type sampled before play	terminal alternative remains physically undecided
State of uncertainty	common belief over type	continuation capability / executable alternatives
Critical event	information revelation / belief splitting	persistent physical loss / commitment
Theory emphasis	value existence, primal/dual HJ characterization, behavioral strategy synthesis	response reduction plus policy-space column generation and oracle-gap certificate
Learning object	numerical approximation of value / behavioral strategy	newly trained high-level game policies with frozen whole-body skills
Primary approximation	value/belief discretization and function approximation	finite response-basis gap δ_B and oracle error ϵ_k

optimization is run. Simulator environments run in parallel for throughput but do not create additional statistical units.

Fair training budget. E1/E2 comparisons match the number of game-actor environment transitions. Forced-response evaluation transitions are reported separately. S-WBC training is not repeated and is not counted toward any method's budget because all methods use the same frozen skills.

Confidence intervals. Every cell reports its count. Intervals are 95% percentile intervals from an environment-cluster bootstrap within each (policy, response) cell ($B = 2000$); for worst-case statistics the maximum over response rows is recomputed inside each replicate with mixture weights held fixed, and reduced-cost intervals re-solve the restricted game in each replicate. Negative reduced costs are reported rather than clipped.

Best-response audit. The global neural oracle error ϵ_k is not directly identifiable from finite runs. We therefore report all oracle seeds, the best held-out value among them, and whether the dual mixture outperforms alternative training distributions.