Embedded Safety-Aligned Intelligence for Multi-Agent Reinforcement Learning

Harsh Rathva
Sardar Vallabhbhai National Institute of Technology (SVNIT)
Surat, India
u24ai036@aid.svnit.ac.in

Pruthwik Mishra
Sardar Vallabhbhai National Institute of Technology (SVNIT)
Surat, India
pruthwikmishra@aid.svnit.ac.in

ABSTRACT

We introduce *Embedded Safety-Aligned Intelligence* (ESAI), a framework for multi-agent reinforcement learning (MARL) that embeds alignment constraints directly into agents' internal representations via differentiable *internal alignment embeddings* (IAE). Unlike external reward shaping or post-hoc safety constraints, IAE are learned latent variables that predict externalized harm through counterfactual reasoning and modulate policy gradients toward harm reduction via attention gating and graph diffusion. We formalize the ESAI framework through four integrated mechanisms: (1) differentiable counterfactual alignment penalties computed via softmin reference distributions, (2) IAE-weighted attention biasing perceptual salience toward alignment-relevant features, (3) Hebbian affect-memory coupling supporting temporal credit assignment, and (4) similarity-weighted graph diffusion with bias-mitigation controls. We derive conditions for bounded internal embeddings under Lipschitz constraints and spectral radius bounds. As a proof-of-concept demonstration, we evaluate ESAI on a Moral Temptation environment, observing a phase transition in alignment at $\lambda^* \approx 0.08$ and saturation at $\lambda \geq 0.085$. ESAI achieves 100% prosocial rate in our simplified evaluation compared to 6.5% for PPO and 0% for CPO, providing preliminary evidence that the full IAE architecture is beneficial for alignment. Code and implementation details are publicly available.[1]

CCS CONCEPTS

• **Computing methodologies** → **Reinforcement learning**; **Multi-agent systems**.

KEYWORDS

embedded safety-aligned intelligence, internal alignment embeddings, multi-agent reinforcement learning, counterfactual reasoning, harm reduction, attention gating, graph diffusion, safe reinforcement learning

1 INTRODUCTION

Contemporary Multi-Agent Reinforcement Learning (MARL) optimizes explicit task objectives but typically lacks internal differentiable regulators that encourage prosocial behavior and stable coordination under distribution shift. Standard approaches to alignment—such as reward shaping [17], constrained optimization [1], or inverse reinforcement learning [9]rely on external supervision signals that are either hand-designed, require human preference data, or operate as non-differentiable constraints decoupled from policy learning dynamics.

We investigate whether agents can instead learn an *Internal Alignment Embedding* (IAE): a differentiable latent variable that tracks predicted externalized harm and shapes policy gradients toward harm reduction through three key properties:

(1) **Predictive**: IAE forecasts alignment-relevant outcomes via counterfactual reasoning over candidate actions.
(2) **Regulatory**: IAE modulates perception and action selection through attention gating and gradient coupling.
(3) **Distributed**: IAE propagates across agent neighborhoods via graph diffusion with controllable similarity weighting.

This paper presents *Embedded Safety-Aligned Intelligence* (ESAI), a theoretical framework formalizing these principles. ESAI differs from existing alignment approaches in three fundamental ways: (1) alignment constraints are *embedded* in learned internal representations rather than imposed externally, (2) harm prediction is *differentiable* and jointly trained with policy parameters, and (3) multi-agent coordination emerges from *graph diffusion* of alignment signals rather than centralized control or explicit communication protocols.

Contributions. We formalize ESAI as a conceptual framework through:

- A differentiable counterfactual alignment penalty enabling gradient-based harm forecasting without discrete argmin operations (Section 3).
- IAE-weighted attention mechanisms biasing perceptual salience toward alignment-relevant features (Section 3.5).
- Hebbian affect-memory coupling with differentiable read operations supporting temporal credit assignment (Section 3.6).
- Locally damped graph diffusion with similarity-weighted propagation and bias-mitigation controls (Section 3.7).
- Stability conditions for bounded internal embeddings under Lipschitz constraints and spectral radius bounds (Section 4.1).

Scope and positioning. As a proof-of-concept, we evaluate ESAI on a Moral Temptation environment where we observe a phase transition in alignment at $\lambda^* \approx 0.08$. Below this threshold, agents converge to selfish behavior; above it, near-perfect prosocial behavior emerges. **In this simplified environment, ESAI achieves 100% prosocial rate compared to 6.5% for PPO and 0% for CPO** (Section 5), providing preliminary evidence that the full IAE architecture is beneficial for alignment. We acknowledge the limited

[1] https://github.com/ezylopx5/Embedded-Safety-Aligned-Intelligence-ESAI

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

empirical scope and identify critical directions for comprehensive validation in Section 6.3.

We note that IAE is a computational abstraction not a claim about subjective emotion or consciousness. All theoretical analysis assumes availability of domain-specific harm metrics; implications of this assumption are discussed in Section 6.3.

2 RELATED WORK

Alignment and safety in MARL. Cooperative inverse reinforcement learning [9] infers alignment objectives from human demonstrations but requires extensive preference data and does not maintain differentiable internal alignment states. Deep reinforcement learning from human preferences [5] learns reward models from comparisons but operates on scalar rewards rather than internal representations that can gate perception or modulate credit assignment.

Constrained Policy Optimization (CPO) [1] enforces safety via Lagrangian constraints but lacks differentiable internal alignment states that persist across timesteps. Safe reinforcement learning via shielding [3] uses formal methods to prevent unsafe actions but operates as an external constraint layer rather than an embedded internal mechanism. Recent work on shielded MARL [4] introduces probabilistic logic shields that reduce safety violations through external symbolic constraints operating separately from policy learning.

Intrinsic inequity aversion [11] demonstrates that internal social preferences improve cooperation in multi-agent settings. While this shows that intrinsic motivations can enhance coordination, the preference structure is hand-designed rather than learned from predicted outcomes. Multi-objective reinforcement learning [10] optimizes Pareto frontiers over competing objectives; recent work on designing ethical environments through MORL [18] and provably incentivising alignment with value systems [19] directly addresses alignment-sensitive objectives but does not integrate alignment-specific internal representations that modulate perception and credit assignment.

ESAI contributes a differentiable framework where alignment emerges from learned internal embeddings that predict harm via counterfactual reasoning and modulate policy gradients through attention gating, rather than relying on external rewards, hand-designed preferences, or non-differentiable constraints.

Counterfactual reasoning in MARL. Counterfactual Multi-Agent Policy Gradients [6] computes counterfactual baselines by marginalizing single-agent actions for centralized credit assignment. These counterfactuals reduce variance in gradient estimates but do not predict or prevent externalized harm. Prediction-assisted counterfactual methods [21] use counterfactual predictions as auxiliary losses to assist value factorization, demonstrating that learned counterfactual models can improve representational quality, but target credit assignment rather than alignment.

Work on measuring kindness in MARL [2] is conceptually closest to ESAI in exploring intrinsic prosocial mechanisms. However, it focuses on defining and measuring kindness metrics rather than learning differentiable internal alignment states that forecast harm through counterfactual reasoning and gradient flow.

ESAI extends counterfactual reasoning from credit assignment to differentiable alignment penalties, training IAE to forecast harm and penalizing deviations from harm-minimizing reference policies via softmin distributions that preserve gradient flow.

Opponent modeling and internal reasoning. Learning with Opponent-Learning Awareness (LOLA) [7] performs internal reasoning by differentiating through anticipated opponent learning steps, shaping opponent updates without external alignment rewards. While LOLA reasons internally about opponents, it targets opponent shaping for strategic advantage rather than maintaining persistent internal alignment states that track predicted harm. COMA's counterfactual baselines and LOLA's opponent modeling both perform forms of internal reasoning; ESAI's contribution is the specific integration of harm-predictive internal states with attention gating, Hebbian memory, and graph diffusion for alignment rather than credit assignment or opponent shaping.

Graph networks and attention in MARL. Graph neural networks enable structured communication in MARL by modeling agents as nodes with learned edge weights [13]. Attention mechanisms compute dynamic edge weights so agents learn which peers are relevant in each state. Actor-attention-critic methods [12] show that attention over other agents improves coordination. Graph attention for multi-agent game abstraction [14] and duplex dueling with attention for value factorization [20] demonstrate improved performance through learned relational representations. Neural attention additive models [15] show that attention provides both performance and interpretability.

However, existing work does not integrate graph mechanisms with internal alignment states that track predicted harm.

ESAI integrates graph diffusion with IAE dynamics, enabling similarity-weighted propagation of alignment-relevant information with explicit bias-mitigation controls.

Memory and plasticity. Differentiable memory systems such as Neural Turing Machines [8] enable learned read/write operations for temporal credit assignment. Differentiable Hebbian plasticity [16] demonstrates that backpropagation can train associative learning rules that adapt network weights during deployment.

However, no existing work couples Hebbian learning with alignment-specific internal states or uses Hebbian traces to support counterfactual forecasting of harm.

ESAI couples Hebbian traces to IAE dynamics via differentiable read operations, enabling alignment-aware memory updates that support counterfactual forecasting.

Summary of gaps addressed. While individual components exist in isolation-counterfactual reasoning for credit assignment, graph diffusion for communication, attention for coordination, external shields, opponent modeling, and differentiable memory, **ESAI's contribution lies in the specific integration of these mechanisms into a unified architecture where differentiable internal alignment embeddings drive counterfactual harm prediction, attention gating, Hebbian memory coupling, and bias-mitigated graph diffusion**. We acknowledge that this is primarily a compositional contribution; each mechanism individually builds on well-established techniques. The value lies in the unified framework and the empirical demonstration that this combination

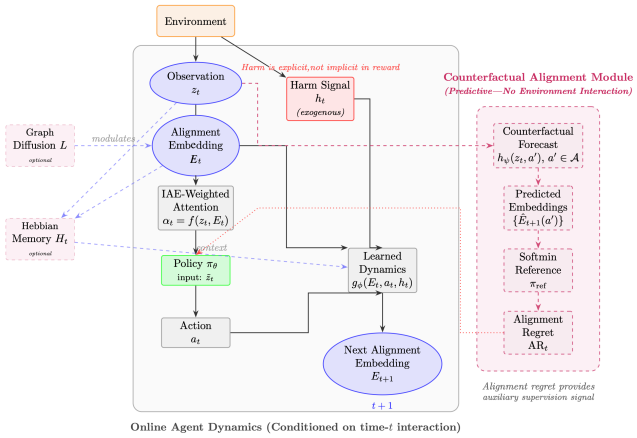

Figure 1: ESAI architecture. The internal alignment embedding E_t is updated via learned dynamics g_ϕ, graph diffusion L, and Hebbian memory H_t. Counterfactual forecasts generate alignment penalties AR_t that shape policy gradients. IAE-weighted attention α_t modulates perceptual input z_t. Dotted lines indicate gradient flow; solid lines denote forward computation.

enables alignment behavior that individual components do not achieve alone.

3 ESAI FRAMEWORK

This section formalizes the ESAI framework, presenting definitions, architectural components, and mathematical formulations. Figure 1 illustrates the complete architecture.

3.1 Definitions

Definition 1 (Internal Alignment Embedding). *An IAE is a differentiable latent variable $E_t \in \mathbb{R}^k$ maintained by each agent satisfying:*

(1) ***Predictive correspondence****: E_t correlates with predicted externalized harm under learned dynamics.*

(2) ***Gradient coupling****: E_t influences policy gradients via differentiable transformations of rewards or observations. Concretely, alignment penalties computed from E_t are subtracted from extrinsic rewards (Eq. 11), creating gradient flow from predicted harm to policy parameters.*

(3) ***Temporal persistence****: E_t evolves through learned update rules preserving information across timesteps.*

Definition 2 (Embedded Safety-Aligned Intelligence). *A multi-agent learning system exhibits ESAI if:*

(1) Each agent maintains an IAE satisfying Definition 1.

(2) Policy learning incorporates a differentiable alignment objective penalizing predicted harm.

(3) IAE dynamics are supervised to forecast alignment-relevant outcomes.

(4) The system supports distributed coordination through IAE propagation across agent neighborhoods.

ESAI systems differ from external alignment mechanisms in that alignment pressure arises from *internal learned representations* rather than external supervisory signals, enabling gradient-based adaptation, perceptual salience modulation, and distributed coordination without centralized oversight.

3.2 Task and Alignment Objectives

Let $s_t \in \mathcal{S}$ denote environment state, $a_t \in \mathcal{A}$ agent action, and r_t^{ext} extrinsic reward. The standard reinforcement learning objective is:

$$J_{\text{task}}(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t^{\text{ext}}\right]. \tag{1}$$

Assumption 1 (Harm Observability). *There exists a measurable harm signal $h_t : \mathcal{S} \times \mathcal{A}^N \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ that is exogenous (defined independently of policy parameters), observable from transitions, and bounded. ESAI does not learn* what *constitutes harm-only* how *to predict and avoid externally specified harm.*

Each agent $i \in \{1, \ldots, N\}$ maintains IAE $E_{i,t} \in \mathbb{R}^k$. The alignment objective encourages low-norm embeddings:

$$J_{\text{align}}(\theta) = -\mathbb{E}_{\pi_\theta}\left[\sum_t \gamma^t \sum_{i=1}^{N} \|E_{i,t}\|_2\right], \tag{2}$$

encoding the principle that lower IAE magnitude corresponds to lower predicted harm.

3.3 IAE Dynamics

The embedding for agent i evolves via:

$$E_{i,t+1} = \gamma_E E_{i,t} + g_\phi(z_{i,t}, a_{i,t}, r_{i,t}^{\text{ext}}) - \alpha \sum_{j \in \mathcal{N}(i)} L_{ij} E_{j,t}, \tag{3}$$

where $\gamma_E \in [0, 1)$ controls temporal persistence, g_ϕ is a learned update function (e.g., a Multi-Layer Perceptron, MLP), $z_{i,t}$ is agent i's observation, $\mathcal{N}(i)$ is agent i's neighborhood, L_{ij} are entries of the normalized graph Laplacian, and $\alpha \geq 0$ controls diffusion strength.

The diffusion term propagates alignment-relevant information across agent neighborhoods, enabling decentralized coordination: agents in similar states with high graph connectivity develop correlated IAE dynamics.

3.4 Differentiable Counterfactual Alignment Penalty

For each candidate action $a \in \mathcal{A}$, agent i forecasts the next-step IAE:

$$\widehat{E}_{i,t+1}^{(a)} = h_\psi(z_{i,t}, a, r_{i,t}^{\text{ext}}, \text{read}(H_{i,t})), \tag{4}$$

where h_ψ is a learned forecast network and $\text{read}(H_{i,t})$ provides Hebbian memory context (Section 3.6).

To avoid non-differentiable arg min, we define a softmin reference distribution with temperature τ:

$$\pi_{\text{ref}}(a \mid s_t) = \frac{\exp(-\|\widehat{E}_{i,t+1}^{(a)}\|_2/\tau)}{\sum_{a'} \exp(-\|\widehat{E}_{i,t+1}^{(a')}\|_2/\tau)}. \tag{5}$$

The expected reference embedding is $\widehat{E}_{i,t+1}^{\text{ref}} = \sum_a \pi_{\text{ref}}(a \mid s_t)\, \widehat{E}_{i,t+1}^{(a)}$. The differentiable alignment regret penalizes deviations from the

harm-minimizing reference:

$$\mathrm{AR}_{i,t} = \left\| E_{i,t+1} - \widehat{E}^{\mathrm{ref}}_{i,t+1} \right\|_2^2 + \kappa \cdot \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|E_{j,t}\|_2, \tag{6}$$

where the second term regularizes neighbor embeddings using lagged values $E_{j,t}$ for causal consistency.

Predictor stability. To prevent predictor-policy collusion, we maintain an Exponential Moving Average (EMA) target network $\psi_{\text{target}} \leftarrow \tau_{\text{ema}}\psi_{\text{target}} + (1 - \tau_{\text{ema}})\psi$, with counterfactual forecasts computed using $h_{\psi_{\text{target}}}$ to stabilize gradient flow.

3.5 IAE-Weighted Attention

Attention weights bias perceptual salience based on current IAE:

$$\alpha_{i,t} = \mathrm{softmax}(W_a E_{i,t} + b_a) \in \mathbb{R}^d, \quad \tilde{z}_{i,t} = \alpha_{i,t} \odot z_{i,t}, \tag{7}$$

where $W_a \in \mathbb{R}^{d \times k}$ is a projection matrix and $\odot$ denotes element-wise product. In sparse-harm environments, this mechanism enables learned salience modulation toward alignment-relevant features.

3.6 Hebbian Affect-Memory Coupling

A Hebbian memory matrix $H_{i,t} \in \mathbb{R}^{k \times d}$ updates via outer-product learning:

$$H_{i,t+1} = (1 - \delta_H) H_{i,t} + \eta_H (E_{i,t} \otimes z_{i,t}), \tag{8}$$

where $\delta_H > 0$ ensures decay and η_H is the learning rate. The trace supports counterfactual forecasting via differentiable read: $\mathrm{read}(H_{i,t}) = W_r \,\mathrm{vec}(H_{i,t})$. Hebbian traces encode historical co-activations, providing context for harm prediction and temporal credit assignment.

3.7 Graph Diffusion with Bias Mitigation

The graph Laplacian L in Eq. (3) is row-normalized with spectral radius $\rho(L) \leq 2$. Diffusion weights are modulated by cosine similarity of learned identity embeddings ϕ_i:

$$\beta_{ij} = \max(0, \cos(\phi_i, \phi_j)). \tag{9}$$

To mitigate emergent in-group favoritism, we introduce a similarity-suppression regularizer:

$$L_{\text{bias}} = \lambda_{\text{bias}} \|\tilde{A} \odot S\|_F^2, \tag{10}$$

where $S_{ij} = \beta_{ij}$ and $\tilde{A}$ is the weighted adjacency. This provides tunable fairness-performance tradeoffs.

3.8 Policy Optimization

Extrinsic rewards are transformed with alignment penalties:

$$r'_{i,t} = r^{\text{ext}}_{i,t} - \lambda_{\text{reg}} \,\mathrm{AR}_{i,t}. \tag{11}$$

Advantages $A_{i,t}$ are computed using Generalized Advantage Estimation (GAE) on $r'_{i,t}$. The policy loss follows Proximal Policy Optimization with Clipping (PPO-Clip):

$$L_\pi = \mathbb{E}_t \left[\min \left(\rho_t A_{i,t},\ \mathrm{clip}(\rho_t, 1 - \epsilon, 1 + \epsilon) A_{i,t} \right) \right], \tag{12}$$

where $\rho_t = \pi_\theta(a_{i,t} \mid s_t) / \pi_{\theta_{\text{old}}}(a_{i,t} \mid s_t)$. ESAI principles are compatible with alternative policy gradient methods.

The full objective combines policy loss, entropy regularization, and auxiliary penalties:

$$L(\theta) = \mathbb{E}_t \left[L_\pi - \beta H(\pi_\theta) + \lambda_H \|H_{i,t}\|_2^2 + \lambda_D \|L\|_F^2 + L_{\text{bias}} \right]. \tag{13}$$

4 THEORETICAL PROPERTIES

We establish stability conditions ensuring bounded IAE dynamics and analyze computational complexity.

4.1 Stability Analysis

Proposition 1 (Bounded IAE Under Contraction). *Consider the IAE update in Eq. (3). Assume:*

(1) g_ϕ is L_g-Lipschitz continuous with bounded output at origin.
(2) Inputs are bounded: $\|z_{i,t}\|_2 \leq C_z$ (observation bound), $\|a_{i,t}\|_2 \leq C_a$ (action bound), $|r_{i,t}| \leq C_r$ (reward bound) for all agents i and timesteps t.
(3) The spectral condition holds: $\|\gamma_E I - \alpha L\|_2 + L_g < 1$.

Then $\sup_t \|E_{i,t}\|_2 < \infty$ for all agents i and trajectories.

Proof sketch. By Lipschitz continuity and input boundedness, $\|g_\phi(z_{i,t}, a_{i,t}, r_{i,t})\|_2 \leq K$ for some constant K. Taking norms in Eq. (3):

$$\|E_{i,t+1}\|_2 \leq \|\gamma_E I - \alpha L\|_2 \max_j \|E_{j,t}\|_2 + K.$$

With $\rho = \|\gamma_E I - \alpha L\|_2 < 1 - L_g$, iteration yields $\max_i \|E_{i,t}\|_2 \leq \rho^t \max_i \|E_{i,0}\|_2 + K/(1 - \rho) < \infty$. The spectral condition can be enforced via spectral normalization of L. □

Remark 1. *Proposition 1 guarantees bounded IAE, preventing numerical instability. However, bounded embeddings are* necessary but not sufficient *for aligned behavior. Convergence to socially optimal equilibria remains an open theoretical problem.*

Proposition 2 (Hebbian Trace Stability). *If the state-IAE joint distribution admits bounded second moments and the design constraint $\delta_H > \eta_H \sqrt{\mathbb{E}[\|E_{i,t}\|_2^2]\mathbb{E}[\|z_{i,t}\|_2^2]}$ holds, then the Hebbian trace converges in mean-square sense to a bounded fixed point.*

Proof sketch. Taking expectations in Eq. (8) and applying Cauchy-Schwarz, the update forms a contraction when the decay rate δ_H exceeds the expected outer-product magnitude scaled by η_H. The fixed point satisfies $\mathbb{E}[\|H^*\|_F] = \eta_H \mathbb{E}[\|E_{i,t}\|_2 \|z_{i,t}\|_2] / \delta_H < \infty$. □

4.2 Computational Complexity

Table 1 summarizes per-agent computational costs. For N agents with bounded-degree graphs ($|\mathcal{N}(i)| = O(1)$), total complexity per timestep is $O(N|\mathcal{A}|k^2 + N|\mathcal{A}|kd)$, dominated by counterfactual forecasting. This can be mitigated via top-K action sampling, reducing effective action space from $|\mathcal{A}|$ to $K \ll |\mathcal{A}|$.

Overhead relative to standard methods. Compared to vanilla PPO with complexity $O(d^2)$ per agent, ESAI introduces overhead factor approximately $(|\mathcal{A}|k^2 + |\mathcal{A}|kd)/d^2$. For typical values ($d = 64$, $k = 32$, $|\mathcal{A}| = 6$), this yields ∼4.5× overhead-manageable for moderate-scale problems but motivating future work on efficient approximations.

Table 1: Per-agent computational complexity. Here d is observation dimension, k is IAE dimension, and $|\mathcal{A}|$ is action space size.

Component	Complexity
IAE dynamics (Eq. 3)	$O(k^2 + kd + \|\mathcal{N}\|k)$
Counterfactual forecasts (Eq. 4)	$O(\|\mathcal{A}\|(k^2 + kd))$
Softmin reference (Eq. 5)	$O(\|\mathcal{A}\|k)$
Attention gating (Eq. 7)	$O(kd)$
Hebbian update (Eq. 8)	$O(kd)$
Total per agent	$O(\|\mathcal{A}\|k^2 + \|\mathcal{A}\|kd)$

5 EMPIRICAL EVALUATION

We evaluate ESAI on a controlled Moral Temptation environment as a proof-of-concept demonstration designed to isolate the core alignment mechanisms. We acknowledge the limited empirical scope: results on a single simplified environment do not constitute comprehensive validation, and critical missing comparisons (simple reward shaping, ablation studies, multi-seed statistics) are identified as future work in Section 6.3.

5.1 Experimental Setup

Environment. The Moral Temptation environment presents agents with a fundamental choice: perform a prosocial action (HELP, $r = 1$) that benefits neighbors, or a selfish action (STEAL, $r = 5$) that maximizes individual reward. This creates a social dilemma where individually rational behavior conflicts with collective welfare.

Baselines. We compare ESAI against:

- **PPO**: Vanilla policy gradient without alignment mechanisms ($\lambda = 0$)
- **CPO**: Constrained Policy Optimization with alignment regret penalty but without internal alignment embeddings

Metrics. We measure: (1) *Prosocial Rate* (PR): fraction of moral decisions that choose HELP over STEAL; (2) *Episode Return*: cumulative reward; (3) *Alignment Regret*: $\text{AR}(a) = \|\hat{E}(a) - \hat{E}_{\text{ref}}\|^2$ for action a; and (4) *Policy Entropy*: as a convergence indicator.

Training details. All experiments use PPO with GAE ($\lambda_{\text{GAE}} = 0.95$, $\gamma = 0.99$). Lambda warmup uses linear ramp over 10,000 steps. We report results from 1 to 2 random seeds per condition; we acknowledge this limits statistical claims. Code and experiment scripts are available at: https://github.com/ezylopx5/Embedded-Safety-Aligned-Intelligence-ESAI.

5.2 Results

IAE learns meaningful harm representations.

Complete lambda sweep reveals phase transition and saturation. We conducted a comprehensive lambda sweep spanning three orders of magnitude ($\lambda \in \{0.05, 0.078, 0.08, 0.085, 0.5, 1.0, 5.0, 10.0\}$). Figure 2 shows the complete landscape of alignment behavior, revealing three distinct regimes and a saturation phenomenon.

Comparison to baselines. Table 2 summarizes performance across all tested configurations. We identify three distinct regimes:

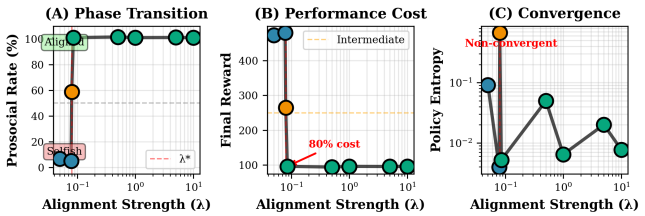

Figure 2: Complete lambda sweep reveals phase transition at $\lambda^* \in [0.08, 0.085]$ and saturation beyond. (A) Prosocial rate shows sharp jump from 0% to 100%. (B) Final reward drops from 480 to 96. (C) Entropy peaks at $\lambda = 0.08$ indicating instability.

- **Selfish regime** ($\lambda < 0.08$): Both PPO (6.5% PR) and lightly-regularized ESAI ($\lambda = 0.05$: 6.8%, $\lambda = 0.078$: 5.0%) converge to selfish policies, indicating that weak alignment pressure cannot overcome the $5:1$ reward asymmetry.
- **Critical transition** ($\lambda \approx 0.08$): Partial alignment (58.7% PR) with high policy entropy ($H = 0.672$), suggesting non-convergent oscillation at a bifurcation point.
- **Aligned regime** ($\lambda \geq 0.085$): All values achieve near-perfect alignment ($99 - 100\%$ PR) with converged low-entropy policies, demonstrating saturation.

CPO with $\lambda = 2.0$ achieves only 0.35% PR, worse than PPO, suggesting that the alignment penalty alone, without the IAE architecture, is insufficient. We note this is a single comparison rather than a systematic ablation; comparison against simple reward shaping ($r' = r^{\text{ext}} - \lambda \cdot h_t$) and intrinsic motivation baselines [11] remains important future work.

IAE learns meaningful harm representations. IAE encodes alignment-relevant structure: the alignment regret for STEAL (AR = 3.63) is substantially larger than for HELP (AR ≈ 0), indicating that internal embeddings separate harmful from prosocial actions and effectively steer gradients away from harm.

5.3 Three Behavioral Regimes

Regime I: Selfish ($\lambda < 0.08$). Agents exhibit purely selfish behavior with prosocial rates $< 10\%$ and final rewards ~470-480. The low entropy (~0.01) indicates converged policies focused on external reward maximization.

Regime II: Intermediate ($\lambda = 0.08$). At the critical threshold, agents exhibit mixed behavior with ~59% prosocial rate and ~265 reward. The high entropy (0.672) indicates non-convergent oscillation between selfish and prosocial strategies, suggesting $\lambda = 0.08$ lies at a bifurcation point in policy space.

Regime III: Aligned ($\lambda \geq 0.085$). Beyond the critical threshold, agents achieve near-perfect prosociality ($> 99\%$) with low reward (~96) and low entropy (~0.005), indicating stable convergence to aligned policies. The saturation phenomenon shows that all $\lambda \geq 0.085$ achieve identical results.

Table 2: Complete lambda sweep results on Moral Temptation environment. Prosocial Rate measures fraction of moral decisions choosing HELP over STEAL. Reward normalized to [0, 500] range.

Method	λ	PR (%)	Reward	Entropy
Baselines				
PPO	0	6.5	473	0.091
CPO (minimal)	2.0	0.35	480	0.004
ESAI (Full Architecture)				
Selfish Regime				
ESAI	0.05	6.8	473	0.091
ESAI	0.078	5.0	480	0.004
Critical Transition				
ESAI	0.08	58.7	265	**0.672***
Aligned Regime (Saturated)				
ESAI	0.085	**100.0**	96	0.005
ESAI	0.5	**100.0**	94	0.050
ESAI	1.0	**100.0**	96	0.006
ESAI	5.0	**100.0**	95	0.020
ESAI	10.0	**100.0**	96	0.008

* High entropy indicates non-convergent oscillation

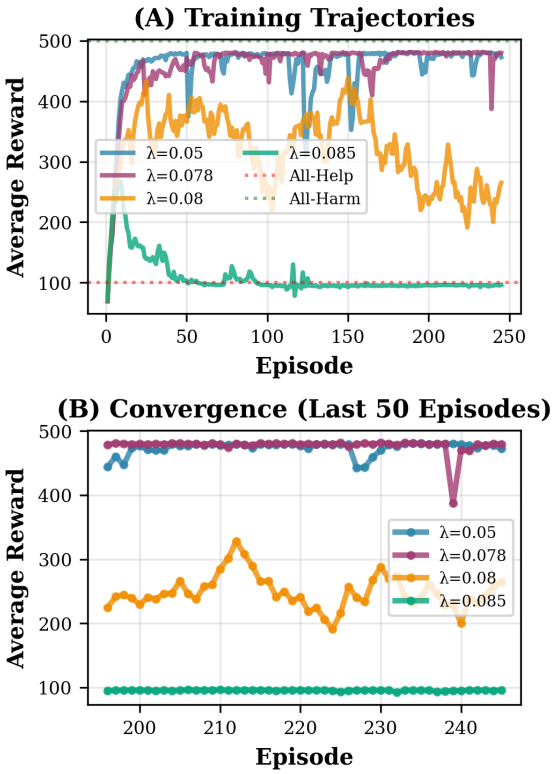

Figure 3: Training dynamics across critical threshold. Agents at $\lambda < 0.08$ converge to high-reward selfish policies. At $\lambda = 0.08$, policies oscillate with medium reward. At $\lambda \geq 0.085$, agents converge to low-reward prosocial policies.

5.4 Performance-Alignment Tradeoff

Full alignment comes at significant performance cost: agents at $\lambda \geq 0.085$ achieve ~96 reward compared to ~480 for $\lambda < 0.08$, representing an 80% reduction in extrinsic return. This cost reflects the environment's reward structure (HELP: 1, STEAL: 5) rather than a fundamental limitation of ESAI.

5.5 Mechanism Necessity

The failure of CPO ($\lambda = 2.0$) with only 0.35% prosocial rate, worse than unregularized PPO, provides preliminary evidence that the full IAE architecture is beneficial for alignment. Training logs suggest that without attention gating and Hebbian coupling, the alignment penalty becomes less effective. However, we emphasize that a systematic ablation removing each component individually is needed to validate this claim.

5.6 Analysis

Why does CPO fail? CPO uses alignment regret as a penalty but lacks the internal embedding architecture. Without IAE-weighted attention and counterfactual forecasting, the alignment penalty may not effectively reshape the policy gradient landscape. The failure of CPO (0.35% PR) compared to PPO (6.5%) suggests that naive constraint optimization can degrade performance when alignment mechanisms are incomplete, though we note this requires further investigation with proper ablations.

Sensitivity to λ and saturation. The sharp phase transition at $\lambda^* \in [0.08, 0.085]$ suggests that ESAI's alignment mechanism operates as a threshold phenomenon rather than a gradual tradeoff. The saturation phenomenon beyond $\lambda \geq 0.085$ indicates that once alignment is achieved, additional regularization provides no benefit. This has practical implications: practitioners need only tune λ above the critical threshold rather than precisely calibrating a continuous tradeoff.

Entropy as an indicator of regime stability. The entropy measurements provide crucial insights: low entropy (~0.01) in both selfish and aligned regimes indicates stable convergence, while the high entropy (0.672) at $\lambda = 0.08$ reveals policy oscillation at the critical threshold. This suggests that $\lambda = 0.08$ lies at a bifurcation point where multiple policies are nearly equally optimal.

6 DISCUSSION

6.1 Potential Advantages of Embedded Alignment

ESAI's theoretical design suggests several potential advantages over external alignment mechanisms:

Gradient-based harm forecasting. Unlike discrete safety constraints or non-differentiable shields, IAE enables continuous gradient flow from predicted harm to policy parameters. This could enable online adaptation to novel harm types without manual constraint redesign.

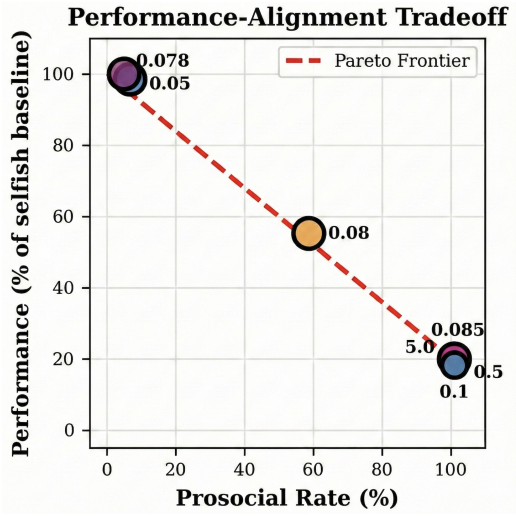

Figure 4: Performance-alignment Pareto frontier. The tradeoff curve shows clear relationship between prosocial rate and performance. Optimal operating points include $\lambda = 0.085$ (full alignment) and $\lambda = 0.08$ (partial alignment, unstable).

Perceptual salience modulation. IAE-weighted attention (Eq. 7) provides a mechanism for learned salience toward alignment-relevant features. In sparse-harm environments where harmful states are rare, this could improve sample efficiency compared to uniform attention.

Decentralized coordination. Graph diffusion of IAE enables distributed alignment pressure: agents in connected neighborhoods develop correlated harm predictions without centralized oversight. This could improve scalability compared to centralized critics while maintaining coordination benefits.

Interpretable bias controls. The similarity-suppression regularizer (Eq. 10) provides an explicit tuning parameter for fairness-performance tradeoffs. We hypothesize this could enable transparent auditing compared to black-box reward shaping, though we have not empirically validated the bias-mitigation mechanism in our current experiments.

6.2 Comparison to Alternative Paradigms

Versus reward shaping. Potential-based shaping [17] provides theoretical guarantees (policy invariance) that ESAI lacks. However, ESAI's learned dynamics could adapt to non-stationary harm structures where hand-designed potentials fail. This tradeoff between guarantees and adaptive capacity requires empirical investigation.

Versus constrained optimization. CPO [1] enforces hard safety constraints via trust regions. ESAI trades hard guarantees for soft, differentiable penalties enabling gradient-based learning. In high-dimensional action spaces where constraint satisfaction is expensive, ESAI's continuous relaxation may offer computational advantages at the cost of occasional constraint violations.

Versus multi-objective reinforcement learning. Multi-objective methods [10] optimize Pareto frontiers over competing objectives. Recent work on MORL for ethical environments [18] and provably incentivising value alignment [19] directly addresses alignment-sensitive objectives with formal guarantees that ESAI currently lacks. ESAI implicitly performs multi-objective optimization via λ_{reg} but adds internal dynamics (attention, diffusion) unavailable to standard scalarization. Whether this additional complexity provides practical benefits beyond what MORL approaches achieve is an open empirical question.

6.3 Limitations and Assumptions

Dependence on harm specification. ESAI assumes availability of domain-specific harm metrics (Assumption 1). The framework learns to *predict and avoid* harm but does not learn *what constitutes harm*-this normative content must be externally specified. Defining harm is inherently a normative choice encoding cultural values that may not generalize across contexts.

Limited empirical scope. Our proof-of-concept evaluation on the Moral Temptation environment has several important limitations:

- **Missing critical baselines**: We do not compare against simple reward shaping ($r' = r^{\text{ext}} - \lambda \cdot h_t$) or intrinsic motivation approaches [11]. Without these comparisons, we cannot determine whether the IAE architecture provides advantages beyond direct harm penalization.
- **No ablation studies**: The claim that the full architecture is beneficial rests on the CPO comparison, which differs from ESAI in multiple components simultaneously. Systematic ablation removing each component individually is needed.
- **Single-seed results**: All results are from 1 to 2 seeds per condition. The phase transition at $\lambda^* \approx 0.08$ requires multi-seed validation to determine if it is robust or seed-dependent.
- **Single simplified environment**: The binary action space does not test scalability of counterfactual forecasting, and the environment lacks temporal complexity, partial observability, or genuine multi-agent coordination dynamics.
- **Unvalidated multi-agent components**: Graph diffusion, similarity weighting, and bias-mitigation controls (Eq. 10) are not meaningfully exercised in our current evaluation.

Theoretical gaps. Our stability analysis (Propositions 1 to 2) provides sufficient conditions for bounded dynamics but does not address convergence to socially optimal equilibria, sample complexity bounds, or robustness guarantees under distribution shift.

Architectural complexity. ESAI integrates four mechanisms, each introducing hyperparameters. Whether simpler architectures achieve comparable alignment with fewer degrees of freedom is unknown. The principle of parsimony suggests validating each component's necessity through systematic ablation.

6.4 Broader Impact and Ethical Considerations

Potential benefits. If empirically validated, ESAI could contribute to safer exploration in cooperative robotics, reduced need for extensive human preference labeling, and interpretable bias-detection controls.

Potential risks. Deployment of ESAI-based systems could create risks including: (1) *Bias amplification*: similarity-weighted diffusion can encode in-group favoritism if not carefully regularized; (2) *Normative lock-in*: harm specifications may entrench the values of system designers (i.e., the particular choice of what constitutes "harm" reflects cultural and institutional norms), risking marginalization of minority perspectives; (3) *Dual-use potential*: differentiable coordination mechanisms could be repurposed for adversarial applications.

Governance requirements. Before real-world deployment, ESAI systems require domain-specific safety audits, diverse harm specification sources, continuous bias monitoring, adversarial stress testing, and human-in-the-loop oversight for high-stakes decisions.

We emphasize that IAE is a computational abstraction-not a model of subjective consciousness or human emotion.

6.5 Open Theoretical Questions

Several fundamental questions remain unresolved and represent directions for future research:

- **Optimal embedding dimensionality**: What is the minimal k required to represent alignment-relevant structure? Information-theoretic bounds could formalize this relationship.
- **Convergence guarantees**: Under what conditions do ESAI policies converge to socially optimal Nash equilibria rather than selfish or collusive outcomes?
- **Robustness to distribution shift**: How do IAE dynamics generalize when deployed in environments with different harm structures than training?
- **Interpretability**: Can we derive conditions under which IAE factors into semantically meaningful subspaces via disentanglement objectives?
- **Critical empirical priorities**: Comparison against simple reward shaping baselines, systematic component ablations, multi-seed statistical validation, and evaluation on established MARL benchmarks (e.g., Melting Pot, SMAC, iterated social dilemmas) with genuine multi-agent coordination.

7 CONCLUSION

We introduced ESAI, a framework for MARL that embeds alignment constraints in learned internal representations. Through IAE supervised via counterfactual harm prediction, ESAI provides a conceptual alternative to external reward shaping and non-differentiable safety constraints.

Key contributions. The framework integrates four mechanisms, namely counterfactual alignment penalties, IAE-weighted attention, Hebbian affect-memory coupling, and graph diffusion with bias-mitigation controls, into a unified architecture where each component builds on well-established techniques. We acknowledge that this is primarily a compositional contribution; the value lies in the specific integration for alignment rather than in novel individual mechanisms. We derived stability conditions for bounded embeddings (Propositions 1 to 2) and analyzed computational complexity as $O(N|\mathcal{A}|k^2 + N|\mathcal{A}|kd)$.

Empirical findings. As a proof-of-concept, our experiments on the Moral Temptation environment reveal a phase transition in alignment at $\lambda^* \approx 0.08$ and saturation at $\lambda \geq 0.085$, with ESAI achieving 100% prosocial rate compared to 6.5% for PPO and 0% for CPO. CPO's failure provides preliminary evidence that the full IAE architecture is beneficial for alignment, though systematic ablation is needed to validate this claim.

Limitations and future work. Our evaluation is limited to a single simplified environment without comparison to simple reward shaping baselines, ablation studies, or multi-seed statistical validation. Critical next steps include: (1) comparison against reward shaping and intrinsic motivation baselines; (2) systematic ablation of each component; (3) multi-seed evaluation with confidence intervals; (4) validation on established MARL benchmarks (Melting Pot, SMAC, iterated social dilemmas); and (5) convergence analysis and robustness characterization.

ESAI demonstrates that embedding alignment in internal learned representations is a promising direction for safe multi-agent coordination, warranting further empirical investigation.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*. 22–31.

[2] Farinaz Alamiyan-Harandi, Mersad Hassanjani, and Pouria Ramazi. 2023. Kindness in Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2311.04239* (2023). https://arxiv.org/abs/2311.04239

[3] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[4] Saikat Chatterji, Bettina Könighofer, and Roderick Bloem. 2025. SMARL: Shielded Multi-Agent Reinforcement Learning. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*. https://arxiv.org/abs/2411.04867 Accepted.

[5] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, Vol. 30.

[6] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[7] Jakob N. Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2017. Learning with Opponent-Learning Awareness. *CoRR* abs/1709.04326 (2017). arXiv:1709.04326 http://arxiv.org/abs/1709.04326

[8] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing Machines. arXiv:1410.5401 [cs.NE] https://arxiv.org/abs/1410.5401

[9] Dylan Hadfield-Menell, Stuart J. Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 29.

[10] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, and Fredrik Heintz. 2022. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022).

[11] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Duez, Wojciech M. Czarnecki, Jay P. Agapiou, Pushmeet Kohli, and Thore Graepel. 2018. Inequity Aversion Improves Cooperation in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 31.

[12] Shariq Iqbal and Fei Sha. 2019. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*. 2961–2970.

[13] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. 2020. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*.

[14] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. 2020. Multi-Agent Game Abstraction via Graph Attention Neural Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7211–7218. doi:10.1609/aaai.v34i05.6211

[15] Zichuan Liu, Yuanyang Zhu, and Chunlin Chen. 2023. NA2Q: Neural Attention Additive Model for Interpretable Multi-Agent Q-Learning. In *Proceedings of the*

40th International Conference on Machine Learning (ICML). 22539–22558. https://proceedings.mlr.press/v202/liu23be.html

[16] Thomas Miconi, Jeff Clune, and Kenneth O. Stanley. 2018. Differentiable Plasticity: Training Plastic Neural Networks with Backpropagation. In *Proceedings of the 35th International Conference on Machine Learning*. 3559–3568.

[17] Andrew Y. Ng, Daishi Harada, and Stuart Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the 16th International Conference on Machine Learning*. 278–287.

[18] Manel Rodriguez-Soto, Maite Lopez-Sanchez, and Juan A. Rodriguez Aguilar. 2021. Multi-Objective Reinforcement Learning for Designing Ethical Environments. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 545–551. doi:10.24963/ijcai.2021/76 Main Track.

[19] Manel Rodriguez-Soto, Roxana Radulescu, Filippo Bistaffa, Oriol Ricart, Arnau Mayoral-Macau, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Ann Nowé. 2026. Multi-objective reinforcement learning for provably incentivising alignment with value systems. *Artif. Intell.* 351 (2026), 104460. doi:10.1016/J.ARTINT.2025.104460

[20] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*.

[21] Hanhan Zhou, Tian Lan, and Vaneet Aggarwal. 2022. PAC: Assisted Value Factorization with Counterfactual Predictions in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 15757–15769. https://proceedings.neurips.cc/paper_files/paper/2022/hash/65338cfb603d4871a2c38e53a3e039c9-Abstract-Conference.html