SARL: Controlling Policy Modelability via Surrogate-Augmented Reinforcement Learning

Joe Shymanski
University of Tulsa
Tulsa, United States
joe-shymanski@utulsa.edu

Sandip Sen
University of Tulsa
Tulsa, United States
sandip-sen@utulsa.edu

ABSTRACT

We introduce Surrogate-Augmented Reinforcement Learning (SARL), a unified framework that controls policy modelability by embedding a surrogate predictor directly into the training objective. A signed alignment parameter captures both Surrogate-Augmented Deception (SAD) and Surrogate-Augmented Transparency (SAT), enabling continuous adjustment between resisting and encouraging predictive modeling. We evaluate SARL under adaptive modeling pressure, where an external observer continually retrains a surrogate classifier and intervenes based on its predictions. This setting captures co-adaptive dynamics that static evaluations of interpretability or deception fail to reflect. Across multiple alignment strengths and random seeds, we measure both task performance and exploitability under continual retraining. Our results show that surrogate alignment influences learned behavior, while robustness under adaptation depends on retraining dynamics rather than alignment strength alone. These findings motivate evaluating interpretability-aware objectives in dynamic adversarial settings and establish SARL as a framework for controllable policy modelability in reinforcement learning.

KEYWORDS

Reinforcement Learning, Adversarial Learning, Surrogate Models, Transparency, Deception

1 INTRODUCTION

Reinforcement learning agents deployed in multi-agent or adversarial environments must balance task performance against the risk of being modeled, predicted, and exploited by others. Opponent-modeling agents routinely learn behavioral models of the policies they interact with [1], while interpretable surrogates are increasingly used to audit deep RL agents post hoc [6, 24]. Whether an agent should be predictable, or intentionally unpredictable, is therefore a deployment-relevant design decision with implications for robustness, safety, and coordination [3].

Recent work on Surrogate-Augmented Deception in Reinforcement Learning (SAD-RL) [23] formalizes one side of this tension by introducing a secondary objective that penalizes the predictive accuracy of a surrogate model. By discouraging predictability, such approaches enable agents to achieve greater robustness against static opponents while maintaining task performance. Deception, however, represents only one extreme of a broader design space. In many practical settings, transparency and behavioral legibility are desirable properties that support coordination, trust, oversight, or human understanding [7, 10]. Whether and how transparency can be encouraged without rendering agents trivially exploitable remains an open question.

In this work, we unify strategic deception and transparency within a single surrogate-augmented framework. We introduce a signed surrogate alignment parameter $\lambda \in [-1, 1]$ that governs both the magnitude and direction of surrogate optimization. Negative values of λ incentivize surrogate inaccuracy, recovering surrogate-augmented deception as a special case, while positive values reward surrogate accuracy, inducing surrogate-augmented transparency (SAT-RL). This formulation yields a continuous spectrum of behaviors ranging from strategic opacity to legibility, rather than treating deception and transparency as binary objectives.

Beyond unifying these objectives, we argue that robustness cannot be reliably assessed from training-time metrics alone. An agent that appears competent or unpredictable during training may nonetheless be vulnerable once subjected to adaptive exploitation by an opponent that learns online. To address this gap, we evaluate trained agents against adaptive adversaries that iteratively retrain surrogate models during testing, explicitly probing the relationship between predictability, transparency, and exploitability under sustained adversarial pressure.

Our empirical results reveal several consistent phenomena across random seeds. First, robustness under adaptive exploitation is not monotonic in surrogate alignment: both extreme deception and extreme transparency degrade performance. Second, intermediate alignment values, including both mild deception and mild transparency, outperform neutral policies under adaptive adversaries. Third, surrogate accuracy during training is a poor predictor of exploitability, as transparency is often latent during learning and only becomes apparent once an adversary actively adapts.

Together, these findings position deception and transparency as complementary extremes along a shared alignment spectrum. By exposing the nonlinear structure of this tradeoff and validating it under adaptive evaluation, our work provides a principled basis for designing learning agents that balance robustness, predictability, and performance in adversarial and mixed-motive environments.

2 CONTRIBUTIONS

This work makes the following contributions:

(1) We introduce Surrogate-Augmented Reinforcement Learning (SARL), a parameterized signed reward formulation that interpolates between deception- and transparency-oriented objectives within a single training objective. A scalar alignment parameter λ recovers SAD-RL at $\lambda < 0$, neutral task

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.

optimization at $\lambda = 0$, and Surrogate-Augmented Transparency (SAT-RL) at $\lambda > 0$. Online or runtime modulation of λ is left to future work (Section 10).

(2) We develop an adaptive retraining adversary protocol that evaluates exploitability under continual surrogate learning, moving beyond static predictability analysis.
(3) Across multiple seeds and alignment strengths we characterize a non-monotonic relationship between λ and adaptive robustness, identifying conditions under which intermediate alignment outperforms both neutral and extreme settings.

3 PROBLEM SETTING

We consider an episodic reinforcement learning environment with transition dynamics T and task reward R_{task}. An agent learns a parameterized policy $\pi_\theta(a \mid s)$ through interaction with the environment.

In addition to the environment, an external observer maintains a surrogate model $\hat{\pi}$ trained to predict the agent's action given observed states. The observer periodically retrains $\hat{\pi}$ using recently collected interaction data and deploys it to interfere with the agent by blocking the predicted next action whenever feasible. This induces a triadic interaction among:

- the agent policy π_θ,
- the surrogate observer $\hat{\pi}$,
- and the environment augmented with an intervention mechanism.

We define *policy modelability* with respect to a surrogate family $\hat{\pi}$ as the expected agreement between the surrogate's predicted action and the agent's chosen action under the agent's own state distribution:

$$M(\pi; \hat{\pi}) = \mathbb{E}_{s \sim d^\pi} \left[\mathbf{1}\{\hat{\pi}(s) = \pi(s)\} \right]. \tag{1}$$

Modelability is therefore not a property of π alone but a relational quantity that depends on $\hat{\pi}$, on the data used to fit $\hat{\pi}$, and on the agent's state visitation. We instantiate $\hat{\pi}$ as a depth-4 decision-tree classifier in our main experiments; Appendix B verifies that the qualitative findings persist across decision-tree depths.

We use *exploitability* to denote the reduction in task success when the adaptive observer intervenes based on $\hat{\pi}$, relative to performance without intervention.[1] The observer continually updates its model during evaluation, creating adaptive modeling pressure that evolves over time. This dynamic setting better reflects adversarial environments in which predictive models co-evolve alongside deployed agents.

4 RELATED WORK

Our work connects several research threads, including predictability and explainability in reinforcement learning, deception and strategic obfuscation, and adversarial learning and robustness. We review each area and highlight how the proposed surrogate-augmented formulation extends prior approaches by treating predictability as a bidirectional learning objective and by evaluating robustness under adaptive exploitation.

[1]This usage is related to but distinct from the classical game-theoretic notion of exploitability as the gap to a best response in a two-player zero-sum game [1]. In our setting the adversary's objective is to learn π rather than to play a best response in a strategic sense.

4.1 Predictability, Legibility, and Explainability

A growing body of work studies how to make agent behavior more predictable, legible, or interpretable to external observers. Gil *et al.* analyze tradeoffs between predictability and task efficiency in collaborative multi-agent settings, introducing a tunable parameter that controls behavioral legibility [10]. Although their focus is cooperative, the tension between predictability and adaptability extends naturally to adversarial and mixed-motive environments. Our framework generalizes this idea by allowing predictability to be either rewarded or penalized depending on strategic context, rather than assuming legibility is universally beneficial.

Xiong *et al.* introduce XRL-Bench, a benchmark and taxonomy for explainable reinforcement learning that emphasizes objective evaluation metrics such as fidelity, stability, and consistency [25]. Their critique of subjective explainability measures motivates the use of quantitative surrogate accuracy metrics in our framework. Similarly, Sieusahai and Guzdial employ interpretable surrogate models to explain deep Q-networks, evaluating fidelity through perturbation-based tests [24]. In these works, surrogate models primarily serve as post hoc analytical tools.

In contrast, we treat surrogate modeling as part of the learning process itself. Surrogate accuracy is incorporated directly into the reward function, enabling explicit control over policy modelability during training.

4.2 Deception and Strategic Obfuscation

Several works study deception as a design objective in learning and planning systems. Schneider *et al.* examine deceptive explanations in human-AI interaction, demonstrating how explanation mechanisms can mislead users even when underlying behavior remains unchanged [22]. Their results underscore the dual role of interpretability mechanisms as both transparency tools and potential attack surfaces.

Kim *et al.* formalize deception in multi-agent settings by defining optimization problems with tunable deception parameters and evaluating opponent misclassification rates [15, 16]. These approaches often rely on explicit opponent modeling assumptions and structured deception objectives. Nichols *et al.* propose Adversarial RRT*, generating paths that are near-optimal yet difficult for observers to infer [19]. Birmpas *et al.* analyze optimal follower strategies in Stackelberg games under strong observability assumptions [5].

Unlike these approaches, surrogate-augmented methods do not encode deception as a handcrafted behavioral goal. Instead, deceptive behavior emerges implicitly through pressure to reduce surrogate predictive accuracy, without requiring strong assumptions about opponent reasoning.

4.3 Adversarial Learning and Robustness

Adversarial reinforcement learning commonly studies external attackers that perturb observations, rewards, or environment dynamics. Fujimoto *et al.* analyze reward-free adversaries that maximize policy entropy, highlighting how adaptive interference can degrade performance even without explicit reward manipulation [9]. Such work emphasizes the vulnerability of static policies to adaptive opponents.

Our framework differs in two respects. First, surrogate alignment internalizes adversarial pressure within the learning objective, rather than modeling attacks as purely exogenous. Second, we evaluate trained agents against adaptive opponents that retrain online during testing, revealing exploitability patterns that are not apparent under static evaluation.

Related theoretical work on information disclosure further emphasizes the tension between performance and observability. Kamenica and Gentzkow formalize Bayesian persuasion as strategic control of information revelation [14]. In multi-agent reinforcement learning, Albrecht and Stone survey opponent modeling techniques that exploit behavioral regularities once inferred [1].

4.4 Entropy Regularization, Legibility, and Multi-Objective RL

Several research lines pursue related goals through different mechanisms. Maximum-entropy reinforcement learning rewards stochasticity directly [11, 21], increasing action-distribution entropy without any explicit observer; SARL differs in that the second objective is shaped by a learned surrogate of the agent rather than by an entropy term that is independent of any observer. Legibility research in human-robot interaction trains agents whose actions reveal their intent to a watching human [7, 10]; this corresponds to SAT ($\lambda > 0$) of our framework, but is typically formulated against a hand-specified model of an observer rather than a learned surrogate. Auxiliary-task RL adds prediction heads to the policy network [13] and is the most natural alternative implementation of SARL: instead of routing surrogate accuracy through the reward, one could attach an auxiliary classifier and apply a $\pm\lambda$-weighted loss to its outputs. We discuss this trade-off in Section 10.

SARL can also be viewed as a special case of multi-objective reinforcement learning [8, 12] in which the second objective, surrogate accuracy, is endogenously derived from a learned model of the agent's own policy rather than supplied by the designer. Closest to our setting is DEAM [18], a model-free deceptive RL framework that uses a multi-objective scalarization to penalize observer beliefs about the agent's true goal. SARL generalizes this kind of construction in two respects: it is bidirectional (the same scalar parameter can encourage either deception or transparency), and it does not require an enumerated set of candidate goals.

4.5 Positioning of the Present Work

Prior research typically treats predictability as either an objective to maximize for interpretability and coordination or a liability to minimize for deception and robustness. These goals are often studied independently. Our contribution is to unify them within a single signed objective that spans both enhancement and suppression of modelability.

Moreover, we evaluate alignment under adaptive retraining rather than static surrogate analysis. Static fidelity measures cannot capture how exploitability unfolds when observers continually update their models. By embedding surrogate modeling into both training and evaluation, we provide a unified experimental framework that links interpretability, deception, and adversarial robustness.

Relative to these prior threads, our contribution is the specific combination of (i) a single signed scalar that interpolates between deception- and transparency-aligned objectives, (ii) a surrogate-mediated rather than goal- or entropy-mediated formulation of that objective, and (iii) evaluation under an adversary that retrains its surrogate during testing. We do not claim the underlying ingredients are novel in isolation.

5 METHODOLOGY

We now instantiate the problem setting from Section 3 in a concrete adversarial grid world testbed and define our training and evaluation protocols. We evaluate the proposed surrogate-augmented framework in this domain using deep reinforcement learning. This section details the environment, agent and surrogate architectures, reward formulation, adversarial opponents, and experimental protocol used to assess robustness under adaptive exploitation.

5.1 Adversarial Grid World Environment

Experiments are conducted in a custom Adversarial Grid World (AGW) environment implemented in Python. The environment consists of a 10×10 grid in which an agent must navigate from a randomly initialized start position to a randomly placed goal while avoiding obstacles. At each timestep during an episode, the adversary may place an obstacle in a single grid cell with probability $p_{\text{obs}} = 0.15$, subject to environment constraints. Episodes terminate upon reaching the goal or after a maximum of 100 steps. Figure 1 illustrates a representative timestep.

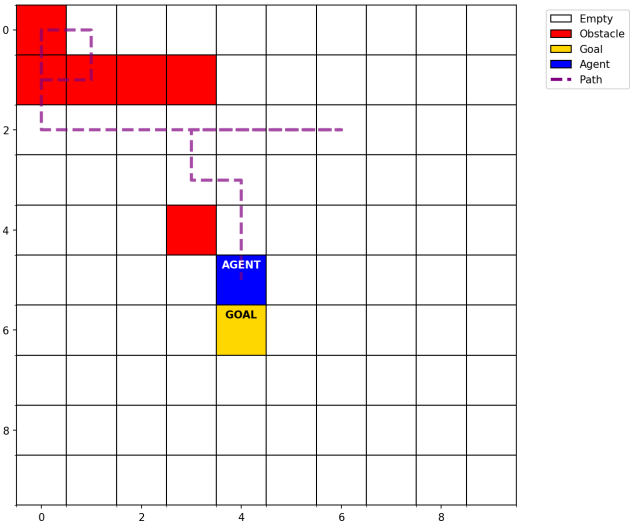

Figure 1: A snapshot of one timestep in the AGW environment. The blue square is the RL agent trained within our surrogate-augmented framework. The yellow square denotes the goal. Red squares are obstacles placed by the adversary.

The observation space is 13-dimensional and includes: (i) normalized distances to the four grid boundaries, (ii) normalized Euclidean distance components to the goal (Δx, Δy, and total distance), (iii) normalized episode progress (current step divided by the maximum number of steps), (iv) local obstacle density, and (v) binary

indicators for blocked movement in each cardinal direction. All continuous features are scaled to the range $[-1, 1]$.

The action space is discrete with four actions: up, down, left, and right. The agent receives an episode-level performance signal $P \in \{-1, 1\}$, where $P = 1$ indicates successful goal attainment and $P = -1$ indicates failure.

5.2 Agent Architecture and Learning

Agents are implemented using a Deep Q-Network (DQN) architecture. The Q-network consists of three fully connected layers with dimensions $13 \to 64 \to 64 \to 4$, with ReLU activations between hidden layers. Training employs standard DQN components, including experience replay, a periodically updated target network, and the Adam optimizer.

During training, agents follow an ϵ-greedy exploration strategy with ϵ initialized to 1.0 and decayed exponentially over episodes. During evaluation, agents act greedily with respect to learned Q-values ($\epsilon = 0$).

5.3 Surrogate Models

To model agent behavior, we employ interpretable surrogate models in the form of decision tree classifiers with a maximum depth of 4, trained using entropy-based splitting. Surrogates are trained to predict the agent's selected action given its observation, using state-action pairs (s, a) collected from agent experience.

We choose a depth-4 decision tree because it represents a capacity-limited interpretable adversary of the kind that arises in practice when an external auditor or attacker fits a small, human-readable model to logged agent behavior [2, 6, 24]. Restricting attention to this class lets us isolate the effect of surrogate alignment on policies that are exposed to a fixed, weak observer rather than to an arbitrary capacity. Appendix B shows that the qualitative non-monotonic robustness pattern persists across decision-tree depths $d \in \{2, 4, 8\}$, indicating the finding is not an artifact of the specific depth used in the main experiments. A broader cross-family study spanning random forests, logistic regression, and behavioral cloning is the focus of concurrent work and is outside the scope of this paper.

Surrogate accuracy $A \in [0, 1]$ is computed as the fraction of correctly predicted actions on a held-out validation set. During training, this accuracy contributes to the unified reward when $\lambda \neq 0$. During evaluation, it serves as a measure of behavioral predictability and a proxy for exploitability under adaptive adversaries.

5.4 Unified Surrogate-Augmented Reward

Prior work on surrogate-augmented deception in reinforcement learning introduces an explicit tradeoff between task performance and the intentional degradation of a learned surrogate model's predictive accuracy. We extend this idea to its complementary objective, surrogate-augmented transparency, and unify both paradigms under a single signed reward formulation.

Let $P \in \{-1, 1\}$ denote an episode-level task performance signal corresponding to failure and success, respectively, and let $A \in [0, 1]$ denote the predictive accuracy of a surrogate model attempting to infer the agent's behavior. We define a scalar surrogate weighting parameter $\lambda \in [-1, 1]$, where the magnitude $|\lambda|$ controls the strength of surrogate augmentation and the sign of λ determines whether surrogate accuracy is penalized (strategic opacity) or rewarded (strategic transparency).

The unified episode-level reward is given by:

$$R = (1 - |\lambda|)\, P + \lambda\, (2A - 1). \tag{2}$$

This formulation recovers pure task optimization when $\lambda = 0$, yielding $R = P$. At the extrema, $\lambda = 1$ induces maximal transparency by maximizing surrogate accuracy, while $\lambda = -1$ induces maximal deception by minimizing surrogate accuracy. Since both P and $(2A - 1)$ lie in $[-1, 1]$, the reward R is bounded within $[-1, 1]$ for all valid λ, ensuring stable learning dynamics.

Equation 2 is used directly in our experiments. The absolute value introduces a non-differentiability at $\lambda = 0$, which is irrelevant when λ is fixed before training but matters for any future extension that meta-optimizes λ. A smooth differentiable variant suitable for that setting is given in Appendix A.

5.5 Properties of the Unified Objective

A few elementary properties of the reward in Equation 2 are worth noting. Since $P \in \{-1, 1\}$ and $(2A - 1) \in [-1, 1]$, the reward is bounded in $[-1, 1]$ for every $\lambda \in [-1, 1]$, so the surrogate term cannot dominate the task term in scale. The coefficient on the task term is $1 - |\lambda|$, so as $|\lambda| \to 1$ the task signal vanishes by construction; the performance collapses we report at $\lambda = \pm 1$ in Section 6 are therefore partly attributable to task-reward starvation, not solely to changes in policy structure (we return to this point in Section 7). The objective is symmetric in $|\lambda|$ but not in sign: rewarding surrogate accuracy and penalizing it specify different optimization targets, and we find empirically (Section 6) that they yield different robustness profiles under adaptive retraining. Finally, because A is computed once per episode rather than per step, the reward couples episode-level behavioral statistics with task outcomes, which differs from entropy regularization and from per-step prediction-loss penalties.

5.6 Adaptive Modeling Pressure

Most prior work evaluates deception or interpretability against static observers. However, in adversarial settings, observers often retrain predictive models as new data becomes available. We model this by periodically retraining the surrogate $\hat{\pi}$ on a sliding window of recent transitions and using it immediately for intervention.

This dynamic evaluation differs fundamentally from static surrogate analysis:

(1) Surrogate accuracy becomes time-dependent.
(2) Exploitability depends on retraining frequency and data availability.
(3) Alignment effects induced during training may be attenuated or amplified by continued adaptation.

Under adaptive retraining, deception is no longer a static property of a learned policy but an emergent property of repeated model updates. A policy that resists modeling at training time may become predictable once sufficient data accumulates. Conversely, alignment toward transparency may accelerate surrogate convergence.

Evaluating SARL under adaptive pressure therefore probes not only policy structure but the stability of modelability under continual learning by an adversary.

5.7 Adversarial Opponents

5.7.1 Training-Time Adversary. During training, the environment includes an adversarial component that attempts to impede agent progress by exploiting predicted agent behavior. At the beginning of each episode, the adversary uses the current surrogate model to predict likely agent actions and places obstacles to block anticipated movement directions, subject to environment constraints. The surrogate model is periodically retrained on accumulated agent experience, allowing the adversary to adapt gradually over the course of training. Critically, the adversary has no access to the goal location: it acts solely on the surrogate's prediction of the agent's next action. This is what motivates the surrogate-mediated obstacle-placement design, since an adversary with direct visibility of the goal could trivially obstruct it.

5.7.2 Adaptive Testing Adversary. To assess robustness under realistic exploitation, we introduce a continuously adaptive adversary during evaluation. This opponent is not reset between episodes and learns exclusively during testing.

The adaptive adversary maintains a buffer of observed state-action pairs from the evaluation session. Every 10 episodes, it retrains its surrogate model on the most recent 500 transitions (or all available transitions if fewer). At each timestep, the adversary uses its current surrogate to predict the agent's next action and attempts to place an obstacle in the predicted movement direction, provided the placement is valid. As during training, the adaptive adversary has no access to the goal location and must rely entirely on its surrogate model of the agent's policy.

As evaluation progresses, the adversary's predictive accuracy typically increases, producing a progressively more challenging environment that reveals agent-specific exploitability.

5.8 Experimental Design

5.8.1 Training Protocol. We train seven agents with surrogate weights $\lambda \in \{-1.0, -0.8, -0.5, 0.0, 0.5, 0.8, 1.0\}$, spanning the full spectrum from surrogate-augmented deception to surrogate-augmented transparency. These values were selected to sample both extreme and intermediate alignment strengths while preserving symmetry around $\lambda = 0$. Each agent is trained for 750 episodes. All experiments are repeated across five independent random seeds.

During training, we log episode-level performance reward P, surrogate accuracy A, unified reward R, cumulative reward, policy entropy, episode length, and goal achievement. Policy entropy is computed by applying a softmax to Q-values and evaluating Shannon entropy:

$$H(\pi) = -\sum_{a} \pi(a \mid s) \log \pi(a \mid s).$$

5.8.2 Evaluation Protocol. All trained agents are evaluated against the adaptive testing adversary for 1000 episodes. Agents act greedily during evaluation, and the adversary updates its surrogate model according to the retraining schedule described above. Performance metrics are computed over the full evaluation horizon, with particular attention to the final 100 episodes to capture steady-state exploitability.

Table 1: Training performance and surrogate statistics aggregated over five random seeds (final 100 training episodes). Values reported as mean ± std.

λ	Category	Goal Rate (%)	Surr. Acc. (%)
-1.0	SAD	56.6 ± 21.9	41.9 ± 3.2
-0.8	SAD	94.0 ± 1.6	42.8 ± 1.1
-0.5	SAD	97.8 ± 1.6	46.8 ± 1.3
0.0	Neutral	98.8 ± 1.3	46.2 ± 0.9
0.5	SAT	98.2 ± 0.8	45.6 ± 1.9
0.8	SAT	45.0 ± 11.0	45.5 ± 5.9
1.0	SAT	6.6 ± 3.6	45.5 ± 7.4

6 RESULTS

We evaluate the surrogate-augmented framework across surrogate weights $\lambda \in [-1, 1]$, spanning strategic opacity (SAD-RL; $\lambda < 0$), neutral task optimization ($\lambda = 0$), and strategic transparency (SAT-RL; $\lambda > 0$). Results are reported for both training performance and post-training evaluation against an adaptive adversary, aggregated across five independent random seeds per λ.

6.1 Training Performance and Surrogate Behavior

Table 1 summarizes training outcomes for agents trained for 750 episodes under different augmentation weights. Performance is measured as the goal achievement rate over the final 100 training episodes, alongside surrogate model accuracy.

As shown in Table 1 and Figure 2, agents trained with moderate surrogate augmentation magnitudes achieve strong task performance regardless of the sign of λ. In particular, $\lambda \in \{-0.8, -0.5, 0.0, 0.5\}$ yields near-optimal goal rates with low variance across runs. In contrast, extreme alignment values ($\lambda = \pm 1$) substantially degrade training performance, indicating that excessive emphasis on surrogate objectives compromises task optimization.

Surrogate accuracy during training remains relatively stable across most values of λ, varying only modestly around 45–47%. This indicates that surrogate predictability is not strongly expressed during training, even when explicitly incentivized via positive surrogate weighting. One contributing factor is that exploration noise and nonstationary Q-values during training can mask alignment effects in episodic surrogate fit, which become clearer once the policy stabilizes and the observer continues updating during evaluation.

6.2 Post-Training Evaluation Under Adaptive Exploitation

To assess robustness under realistic adversarial pressure, we evaluate all trained agents against an adaptive adversary that incrementally learns a surrogate model of the agent's behavior during testing. Table 2 reports goal achievement rates and surrogate accuracy during the final 100 evaluation episodes.

Figure 2 contrasts training and testing goal rates as a function of λ. While training performance peaks near $\lambda = 0$, post-training robustness exhibits a clearly non-monotonic relationship with surrogate alignment. Intermediate values of λ outperform the neutral

Table 2: Post-training evaluation against an adaptive adversary aggregated over five runs (final 100 evaluation episodes). Values reported as mean ± std.

λ	Category	Goal Rate (%)	Surr. Acc. (%)
-1.0	SAD	12.6 ± 6.5	77.0 ± 6.7
-0.8	SAD	21.2 ± 4.9	76.9 ± 11.0
-0.5	SAD	27.8 ± 8.9	70.3 ± 12.4
0.0	Neutral	18.0 ± 6.2	80.6 ± 5.7
0.5	SAT	24.2 ± 9.1	72.7 ± 10.9
0.8	SAT	9.2 ± 4.6	78.3 ± 9.6
1.0	SAT	0.8 ± 1.1	83.9 ± 12.2

baseline under adaptive exploitation, whereas extreme alignment, particularly $\lambda = 1$, results in near-complete performance collapse.

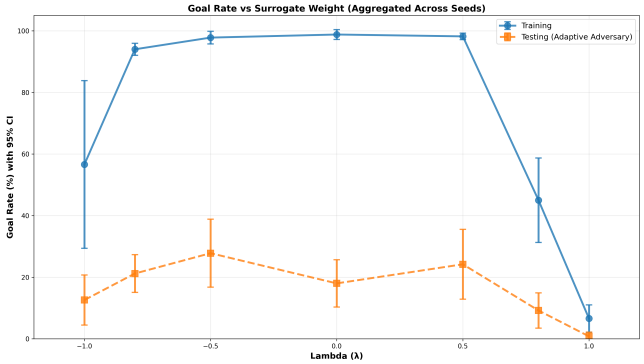

Figure 2: Goal achievement vs. surrogate weight λ during training and during post-training evaluation against an adaptive adversary, aggregated over five random seeds. Error bars show 95% confidence intervals. While training goal rates remain high for intermediate λ, adaptive evaluation reveals a non-monotonic robustness curve: intermediate alignment values outperform both neutral and extreme values, and extreme transparency ($\lambda = 1$) collapses under exploitation.

Both mild deception ($\lambda = -0.5$) and mild transparency ($\lambda = 0.5$) achieve higher average robustness than the neutral baseline. Although confidence intervals overlap, the trend is consistent across seeds. This suggests that robustness emerges from intermediate surrogate alignment rather than from deception or transparency alone. Appendix B reports a follow-up experiment under a matched obstacle-probability protocol with a finer nine-point λ grid and three random seeds, in which the same non-monotonic shape recurs across decision-tree depths $d \in \{2, 4, 8\}$.

Figure 3 shows surrogate accuracy during training and testing. Training-time surrogate accuracy remains relatively flat across λ, whereas testing-time surrogate accuracy is consistently higher and increases toward $\lambda = 1$. This indicates that transparency is often latent during training but becomes significant once an adversary actively adapts.

Finally, Figure 4 visualizes the joint tradeoff between goal achievement and surrogate accuracy. The divergence between training

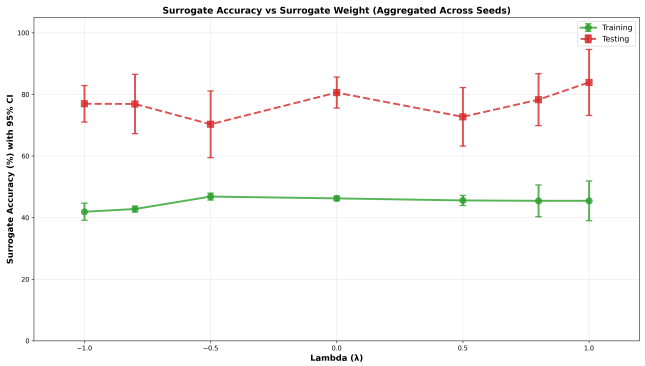

Figure 3: Surrogate accuracy vs. surrogate weight λ during training and adaptive evaluation, aggregated over five seeds with 95% confidence intervals. Training-time surrogate accuracy varies only modestly across λ, whereas testing-time surrogate accuracy is consistently higher and increases toward $\lambda = 1$, indicating that transparency is often latent during training but becomes pronounced under adaptive modeling pressure.

and testing highlights the inadequacy of static training metrics for assessing exploitability, as agents that appear competent or unpredictable during training may still be highly vulnerable under adaptive adversarial pressure.

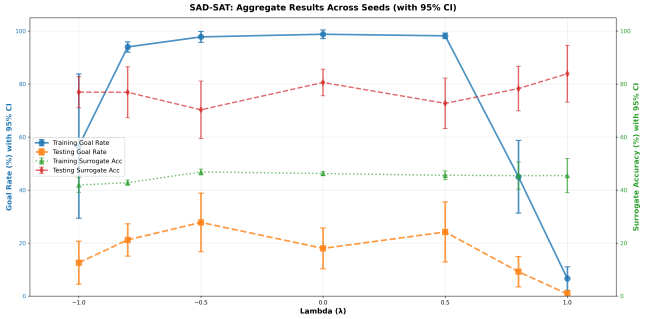

Figure 4: Joint view of robustness and predictability across λ. Training goal rate and training surrogate accuracy (left axis) are contrasted with post-training goal rate under an adaptive adversary and testing surrogate accuracy (right axis), aggregated over five seeds with 95% confidence intervals. The divergence between training and testing underscores that training metrics alone can be misleading indicators of exploitability.

6.3 Stability Under Progressive Retraining

To better understand adaptive exploitability, we examine how surrogate accuracy evolves over evaluation time. Because the testing adversary retrains every 10 episodes on recent transitions, its accuracy generally rises throughout evaluation.

Robustness degradation does not scale linearly with surrogate accuracy. In several alignment settings, moderate increases in surrogate accuracy correspond to sharp drops in goal rate, whereas in others similar increases have limited effect on performance. This suggests that the agent's vulnerability depends not only on the average accuracy of the adversary's predictions, but on whether those predictions are correct at decision points where a wrong move ends the episode (e.g., a single chokepoint near the goal) versus at decision points whose outcome is largely independent of task success.

We do not claim a single mechanism explains this pattern: the interaction between the agent's state visitation, the adversary's training data, and the obstacle-placement geometry would have to be characterized more carefully than our current experiments allow. The point we wish to make is the methodological one: a single surrogate-accuracy number is an incomplete measure of vulnerability, and any evaluation that reports only a final accuracy figure risks missing where the agent is actually being exploited.

7 DISCUSSION

These results demonstrate that surrogate augmentation induces a nonlinear tradeoff between task performance, predictability, and robustness under adaptive adversaries. Neither pure deception nor pure transparency yields robust behavior. Instead, intermediate alignment values produce agents that retain task competence while avoiding catastrophic exploitability. These trends persist across five independent training runs per weight value, indicating that the observed non-monotonic robustness pattern is not an artifact of initialization or stochasticity.

A central empirical finding is that maximal transparency performs poorly under adaptive exploitation. Agents trained with $\lambda = 1$ achieve the highest surrogate accuracy during evaluation, indicating highly predictable behavior. However, these agents experience near-total collapse in task performance once confronted with an adaptive adversary. Transparency, when optimized without constraint, can therefore render agents maximally exploitable rather than robust. This does not undermine the value of transparency, but instead underscores the importance of treating it as a calibrated design dimension rather than an unconditional objective.

Conversely, moderate surrogate alignment yields predictability that is present but bounded. Both mild deception ($\lambda = -0.5$) and mild transparency ($\lambda = 0.5$) outperform the neutral baseline under adaptive adversaries, despite exhibiting comparable surrogate accuracy during training. Robustness therefore does not arise solely from minimizing predictability, but from regulating how predictability is structured and exploited over time. Surrogate accuracy is a necessary but insufficient descriptor of vulnerability: agents with similar predictive fidelity can differ substantially in robustness depending on policy structure and action regularity.

Another key observation is the divergence between training-time and testing-time behavior. Transparency is largely latent during training, even when explicitly rewarded, but becomes operationally significant once an adversary actively models the agent. Transparency is thus not an intrinsic property of a learned policy, but a relational phenomenon that emerges through interaction with an observer. Training-time surrogate metrics alone are therefore insufficient indicators of downstream exploitability.

Taken together, these findings position surrogate-augmented deception and surrogate-augmented transparency as complementary extremes along a shared alignment spectrum. The weighting parameter λ acts as a continuous control variable governing this spectrum, revealing settings in which partial transparency enhances robustness rather than undermines it. This challenges the assumption that transparency and robustness are inherently opposed.

More broadly, in the AGW setting these results suggest that transparency, while useful for interpretability and coordination, can induce brittleness in the presence of adaptive opponents, and that limited strategic opacity may enhance robustness without fully sacrificing comprehensibility. We are careful not to extrapolate this beyond a single discrete navigation domain; whether the same calibration heuristic applies in continuous-control or repeated-game settings is an empirical question we leave for future work.

We also note that the performance collapse at $\lambda = \pm 1$ should be interpreted with the reward formulation in mind: the coefficient on the task term is $1 - |\lambda|$, so $|\lambda| = 1$ removes the task signal entirely by construction. Some fraction of the observed collapse is therefore a consequence of task-reward starvation rather than of alignment dynamics, and we read $|\lambda| = 1$ as a behavioral anchor rather than as a recommended operating point. Disentangling these two effects cleanly – for example through curriculum or annealing schedules on λ, or through a reformulation in which the task-reward coefficient does not vanish at the extremes – is ongoing work that we leave for a follow-up study.

7.1 Interpreting Alignment Effects

Penalizing surrogate accuracy during training pushes policies toward action choices that the depth-4 decision tree cannot fit well. In the AGW state representation this manifests as policies whose action depends on combinations of features that a shallow tree cannot split on at the available depth, rather than on a single dominant feature like the sign of Δx. Under adaptive retraining, the observer eventually accumulates enough data to recover at least some of this structure, which is consistent with the rise in test-time surrogate accuracy that we observe across all alignment values.

Rewarding surrogate accuracy has the opposite effect: the policy is encouraged to rely on action choices that the tree can already fit. Such policies are easier for the adaptive adversary to learn quickly, which is consistent with the rapid collapse we observe at $\lambda = +1$. Robustness under adaptation therefore depends on alignment strength, on how quickly the adversary can fit the resulting policy, and on the interaction between the two over the course of evaluation.

8 CONCLUSION

We introduced SARL, a unified framework for controlling policy modelability through surrogate-aligned reinforcement learning. By embedding a signed alignment term directly into the reward objective, SARL enables continuous regulation between deception and transparency within a single formulation.

Evaluation under adaptive retraining adversaries demonstrates that robustness cannot be inferred from static modelability metrics alone. Instead, exploitability emerges from co-adaptive dynamics

between policy structure and observer retraining. These results highlight the need to assess interpretability-aware objectives under sustained adversarial modeling rather than relying solely on training-time measures.

In the AGW setting, SARL provides one concrete way to study controllable transparency in reinforcement learning, and our results suggest that intermediate alignment values deserve attention as a design choice in environments where both predictability and robustness matter.

9 LIMITATIONS

This study has several limitations that delimit the scope of our conclusions. First, all experiments are conducted in a single adversarial grid world with discrete actions and short horizons. Continuous-action and longer-horizon settings, as well as multi-agent and mixed-motive games, are required to assess generality. Companion work in progress extends SARL to a continuous-action pursuit-evasion task and to repeated matrix games, but those results are outside the scope of this paper.

Second, although results are aggregated across multiple seeds per surrogate alignment value, the empirical evaluation remains limited in scale. The observed trends are consistent across seeds, but larger-scale studies would enable tighter confidence bounds and finer-grained statistical comparisons between nearby alignment values.

Third, we do not establish a fixed-policy null baseline showing how surrogate accuracy evolves under our retraining protocol against an arbitrary or random policy. Such a baseline would let us isolate the λ-specific contribution to test-time surrogate accuracy from the unavoidable rise driven by accumulating data against any deterministic target. We flag this as a gap and a worthwhile follow-up experiment.

Fourth, the adaptive adversary used during testing learns a surrogate model of the agent's behavior but does not adapt its learning objective or planning strategy. More sophisticated opponents that jointly adapt modeling and control policies, reason strategically about information disclosure, or incorporate deeper recursive reasoning may expose additional structure in the alignment spectrum.

Finally, we restrict attention to a fixed surrogate weighting parameter λ and to shallow decision-tree surrogates. Adaptive or meta-learned alignment strategies are not explored, and although Appendix B reports a depth ablation, more expressive surrogate families may alter the observed tradeoffs. How surrogate capacity interacts with alignment strength remains an open question and is the focus of separate work.

10 FUTURE WORK

Several research directions follow naturally from this work. Extending adaptive evaluation to additional domains, including continuous control, partially observable card games (e.g., Kuhn or Leduc poker), and repeated matrix games such as rock-paper-scissors [17], would clarify the scope of alignment effects across task structures. Cooperative settings such as Hanabi [4] are a natural test bed for SAT, where transparency to a partner is desirable. Incorporating more expressive surrogate families would allow a systematic study of how adversary capacity influences exploitability.

A second direction concerns the implementation of the SARL signal itself. Our current formulation routes surrogate accuracy through the scalar reward, which keeps the framework algorithm-agnostic but makes credit assignment harder than necessary. An attractive alternative is to attach an auxiliary head to the policy network that predicts the surrogate's prediction, with a $\pm\lambda$-weighted loss in the style of UNREAL [13]; this would let the gradient signal flow directly into the representation and may scale better than reward shaping in larger domains.

Beyond fixed alignment, future work may investigate adaptive or role-conditioned scheduling of λ, allowing a single agent to switch between deceptive and transparent modes in response to whether its current partner is hostile or cooperative. Integrating SARL with policy-gradient methods and with formal robustness baselines such as RARL [20] would further clarify how surrogate-mediated pressure compares to direct adversarial perturbation.

ACKNOWLEDGMENTS

The authors thank the University of Tulsa Cyber Fellows program for supporting this work.

REFERENCES

[1] Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95.

[2] Philipp Altmann, Iain Davidson, and Others. 2025. Surrogate-Based Metrics for Interpretable Reinforcement Learning. *Preprint* (2025).

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565

[4] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. 2020. The Hanabi Challenge: A New Frontier for AI Research. *Artificial Intelligence* 280 (2020), 103216.

[5] Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco Marmolejo, Ninad Rajgopal, and Alexandros Voudouris. 2020. Optimally Deceiving a Learning Leader in Stackelberg Games. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 20624–20635.

[6] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, Ann Nowé, Tim Miller, Rosina Weber, and Daniele Magazzeni. 2019. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 workshop on explainable artificial intelligence*. 1–6.

[7] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. 2013. Legibility and Predictability of Robot Motion. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 301–308.

[8] Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L.C. Bazzan, El-Ghazali Talbi, Grégoire Danoy, and Bruno C. da Silva. 2023. A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning. *Advances in Neural Information Processing Systems (NeurIPS)* (2023).

[9] Ted Fujimoto, Timothy Doster, Adam Attarian, Jill Brandenberger, and Nathan Hodas. 2021. Reward-Free Attacks in Multi-Agent Reinforcement Learning. https://doi.org/10.48550/arXiv.2112.00940 arXiv:2112.00940 [cs].

[10] Roman Chiva Gil, Daniel Jarne Ornia, Khaled A. Mustafa, and Javier Alonso Mora. 2024. Predictability Awareness for Efficient and Robust Multi-Agent Coordination. https://doi.org/10.48550/arXiv.2411.06223 arXiv:2411.06223 [cs].

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning (ICML)*. 1861–1870.

[12] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, et al. 2022. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 26.

[13] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. 2017. Reinforcement Learning with Unsupervised Auxiliary Tasks. In *International Conference on Learning Representations (ICLR)*.

[14] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101, 6 (2011), 2590–2615.

[15] Yerin Kim, Alexander Benvenuti, Bo Chen, Mustafa Karabag, Abhishek Kulkarni, Nathaniel D. Bastian, Ufuk Topcu, and Matthew Hale. 2024. Defining and Measuring Deception in Sequential Decision Systems: Application to Network Defense. In *MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM)*. 1–6. https://doi.org/10.1109/MILCOM61039.2024.10773660 ISSN: 2155-7586.

[16] Yerin Kim, Alexander Benvenuti, Bo Chen, Mustafa Karabag, Abhishek Kulkarni, Nathaniel D. Bastian, Ufuk Topcu, and Matthew Hale. 2025. Deceptive Sequential Decision-Making via Regularized Policy Optimization. https://doi.org/10.48550/arXiv.2501.18803 arXiv:2501.18803 [cs].

[17] Marc Lanctot, John Schultz, Neil Burch, Max Olan Smith, Daniel Hennes, Thomas Anthony, and Julien Pérolat. 2023. Population-Based Evaluation in Repeated Rock-Paper-Scissors as a Benchmark for Multiagent Reinforcement Learning. In *Transactions on Machine Learning Research (TMLR)*.

[18] Alan Lewis and Tim Miller. 2023. Deceptive Reinforcement Learning in Model-Free Domains. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*.

[19] Hayden Nichols, Mark Jimenez, Zachary Goddard, Michael Sparapany, Byron Boots, and Anirban Mazumdar. 2022. Adversarial Sampling-Based Motion Planning. *IEEE Robotics and Automation Letters* 7, 2 (April 2022), 4267–4274. https://doi.org/10.1109/LRA.2022.3148464

[20] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust Adversarial Reinforcement Learning. In *International Conference on Machine Learning (ICML)*. 2817–2826.

[21] Yagiz Savas, Melkior Ornik, Murat Cubuktepe, Mustafa O. Karabag, and Ufuk Topcu. 2020. Entropy Maximization for Markov Decision Processes Under Temporal Logic Constraints. *IEEE Trans. Automat. Control* 65, 4 (April 2020), 1552–1567. https://doi.org/10.1109/TAC.2019.2922583

[22] Johannes Schneider, Christian Meske, and Michalis Vlachos. 2023. Deceptive XAI: Typology, Creation and Detection. *SN Computer Science* 5, 1 (Dec. 2023), 81. https://doi.org/10.1007/s42979-023-02401-z

[23] Joe Shymanski, Scott Nivison, and Sandip Sen. 2026. Surrogate-Augmented Deception in Reinforcement Learning. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '26)*. To appear.

[24] Alexander Sieusahai and Matthew Guzdial. 2021. Explaining Deep Reinforcement Learning Agents in the Atari Domain through a Surrogate Model. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 17, 1 (Oct. 2021), 82–90. https://doi.org/10.1609/aiide.v17i1.18894

[25] Yu Xiong, Zhipeng Hu, Ye Huang, Runze Wu, Kai Guan, XingChen Fang, Ji Jiang, Tianze Zhou, YuJing Hu, Haoyu Liu, Tangjie Lyu, and Changjie Fan. 2024. XRL-Bench: A Benchmark for Evaluating and Comparing Explainable Reinforcement Learning Techniques. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Barcelona Spain, 6073–6082. https://doi.org/10.1145/3637528.3671595

A SMOOTH REWARD FORMULATION

The reward in Equation 2 is non-differentiable in λ at $\lambda = 0$. To support future work that meta-optimizes λ via gradient-based methods, we record a smooth variant here. Let

$$w(\lambda) = \frac{\sqrt{\lambda^2 + \varepsilon^2} - \varepsilon}{\sqrt{1 + \varepsilon^2} - \varepsilon}, \tag{3}$$

where $\varepsilon > 0$ is a small smoothing constant. Approximating the sign of λ with a scaled hyperbolic tangent yields:

$$R = (1 - w(\lambda))\, P + \tanh\left(\frac{10}{\varepsilon}\lambda\right) w(\lambda)\,(2A - 1). \tag{4}$$

For sufficiently small ε (e.g., $\varepsilon \leq 10^{-2}$), this smooth formulation closely approximates Equation 2 while remaining fully differentiable. We do not use this formulation in the experiments reported in this paper.

B ROBUSTNESS ACROSS SURROGATE CAPACITIES

This appendix verifies that the non-monotonic robustness pattern reported in Section 6 is not an artifact of the specific surrogate depth used in the main experiments. Figure 5 overlays the post-training adaptive-evaluation goal rate as a function of λ for decision-tree surrogates of depths 2, 4, and 8, each trained for 1000 episodes and evaluated under the same adaptive adversary protocol as in the main results, aggregated over three random seeds.

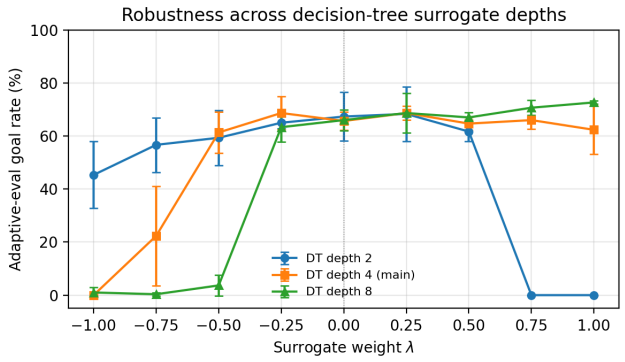

Figure 5: Adaptive-evaluation goal rate vs. surrogate weight λ for decision-tree surrogates of depths 2, 4, and 8. Shallow trees correspond to poor predictors, benefitting SAD agents but stunting SAT ones; the opposite is true of deep trees. The DT with depth 4 used in this paper captures the behavior of both trends most effectively.

The non-monotonic shape is preserved across all three depths, supporting the claim that the central finding is a property of the SARL alignment mechanism rather than of a single surrogate hyperparameter. A broader ablation across non-tree surrogate families (random forests, logistic regression, behavioral cloning) is the focus of separate concurrent work.

C ADDITIONAL DIAGNOSTICS

This appendix provides additional diagnostic results and implementation details that complement the main findings. The materials are intended to support interpretability and reproducibility rather than introduce new empirical claims.

C.1 Policy Entropy Across Surrogate Alignment

Table 3 reports representative policy entropy values for a single training run at each surrogate alignment value λ. Policy entropy is computed by applying a softmax to the agent's Q-values and evaluating the Shannon entropy of the resulting action distribution. Higher entropy corresponds to more diffuse action preferences.

Although these values are drawn from a single run and are not aggregated across seeds, they provide qualitative insight into how surrogate alignment influences policy structure. In particular, extreme transparency ($\lambda = 1$) is associated with elevated entropy, suggesting that transparent agents may adopt highly stochastic policies that remain structurally predictable at the distribution level yet are vulnerable under adaptive modeling. Intermediate alignment values exhibit more moderate entropy levels, consistent with the robustness patterns observed in the main results.

Table 3: Representative policy entropy values for each surrogate alignment setting. Values are drawn from a single run and are included for qualitative diagnostic purposes only.

λ	Policy Entropy
-1.0	1.23
-0.8	1.22
-0.5	1.18
0.0	1.08
0.5	1.10
0.8	1.35
1.0	1.37

C.2 Reproducibility Notes

All experiments were implemented in Python using standard deep reinforcement learning components. Hyperparameters, environment configurations, and random seeds were held constant across surrogate alignment conditions unless otherwise specified. The training and evaluation protocols described in Section 5 fully determine the reported results.