



**MONASH**  
University

**MONASH**  
BUSINESS  
SCHOOL

Department of Economics, University of Warwick  
Monash Business School, Monash University

as part of  
Monash Warwick Alliance

**Learning Correlated Equilibrium Via Neural Network Regret  
Minimisation**

Khushi Sampat

**Warwick-Monash Economics Student Papers**

March 2026

No: 2026/99

ISSN 2754-3129 (Online)

The Warwick Monash Economics Student Papers (WM-ESP) gather the best Undergraduate and Masters dissertations by Economics students from the University of Warwick and Monash University. This bi-annual paper series showcases research undertaken by our students on a varied range of topics. Papers range in length from 5,000 to 8,000 words depending on whether the student is an undergraduate or postgraduate, and the university they attend. The papers included in the series are carefully selected based on their quality and originality. WM-ESP aims to disseminate research in Economics as well as acknowledge the students for their exemplary work, contributing to the research environment in both departments.

**Recommended citation:** Sampat, K. (2026). Learning Correlated Equilibrium Via Neural Network Regret Minimisation. *Warwick Monash Economics Student Papers* 2026/99.

**WM-ESP Editorial Board**

Sascha O. Becker (University of Warwick)

Mark Crosby (Monash University)

Cecilia T. Lanata-Briones (University of Warwick)

Gordon Leslie (Monash University)

Thomas Martin (University of Warwick)

Vinod Mishra (Monash University)

Jeremy Smith (University of Warwick)

Natalia Zinovyeva (University of Warwick)

If you want to get in touch with the Editorial Board, please email [w-mesp@warwick.ac.uk](mailto:w-mesp@warwick.ac.uk)

# Learning Correlated Equilibrium Via Neural Network Regret Minimisation

Khushi Sampat\*

September 2025

## Abstract

This paper studies how decentralised neural agents trained by regret minimisation learn equilibrium behaviour in static games and whether such learning can be extended beyond Nash equilibria. The analysis proceeds in two parts. The first chapter examines equilibrium selection in coordination games with multiple Nash equilibria. Building on recent evidence that neural agents trained across large distributions of games systematically favour risk-dominant equilibria, the chapter introduces a structured pre-training curriculum designed to instil a bias toward payoff-dominant outcomes in Stag Hunt environments. While pre-training successfully induces efficient coordination in these games, the results show that this bias is rapidly eroded under subsequent adversarial training on heterogeneous games, where play reverts to mixed or risk-sensitive equilibria. The second chapter investigates whether decentralised learners can acquire correlated equilibrium behaviour when coordination requires conditioning on private signals. Initial experiments demonstrate that standard personal regret objectives lead agents to ignore mediator signals and converge to unconditional Nash strategies. This limitation is overcome by replacing personal regret with a squared obedience (swap) regret objective. Under this modified objective, neural agents successfully learn signal-contingent behaviour and generalise correlated equilibrium strategies to unseen coordination games. Together, the findings clarify the capabilities and limitations of regret-based learning as a mechanism for equilibrium formation in strategic environments.

**JEL Classification:** C72, C63, C73, D83, C61

**Keywords:** Correlated Equilibrium, Regret Minimisation, Deep Reinforcement Learning, Neural Networks, Game Theory

---

\*khushisampat@gmail.com

Appendix: [git@github.com:khushi228/Neural-Network-Regret-Minimisation.git](https://github.com/khushi228/Neural-Network-Regret-Minimisation.git)

I am grateful to Dr. Daniele Condorelli for his continued guidance and valuable advice throughout the year. I also extend my sincere thanks to Massimiliano Furlan for meeting with me regularly, for patiently explaining key concepts, and for generously sharing resources that greatly supported the development of this paper.

# 1 Introduction

This dissertation explores how neural networks trained by regret minimisation learn to play equilibria in static games. The analysis is divided into two chapters. The first investigates whether the bias towards risk-dominant Nash equilibria identified in recent work (Condorelli and Furlan 2025) persists when agents are exposed to a structured pre-training curriculum designed to instil a preference for payoff-dominant outcomes. The second extends the framework to correlated equilibria, asking whether decentralised agents can learn to condition their play on private signals and coordinate on outcomes that are unattainable in the Nash correspondence.

The two chapters address related aspects of equilibrium formation. The first situates learning within environments characterised by multiple Nash equilibria and asks whether inductive biases can be shifted towards efficiency. The second broadens the scope to coordination devices, showing how changes in objectives and information structure are required for agents to sustain play at correlated equilibria. Together they provide a unified account of how decentralised learning both reinforces existing conventions and, under the right conditions, enables richer forms of strategic behaviour.

While a pre-training curriculum can transiently instil a preference for payoff-dominant outcomes in Stag Hunt games, this bias is rapidly eroded under subsequent adversarial training, where play reverts to mixed or risk-sensitive equilibria. Chapter Two demonstrates that minimising personal regret is insufficient for learning correlated equilibria, as agents learn to ignore mediator signals in favour of unconditional Nash strategies. This failure is overcome by adopting a squared obedience (swap) regret objective, which successfully produces agents

that generalise signal-contingent play to unseen coordination games with high fidelity.

The rest of the dissertation is organised as follows: Chapter One presents the analysis of equilibrium selection under pre-training, and Chapter Two develops the framework for learning correlated equilibria. A final section concludes.

## 2 Chapter One

Nash equilibrium is a central solution concept in game theory. Yet in games with multiple equilibria it leaves unresolved which outcome should be expected. The Stag Hunt is a canonical example, where agents choose between a safe but inefficient strategy and a risky strategy that yields higher payoffs if coordination succeeds. Harsanyi and Selten (1988) propose the risk-dominant equilibrium as the natural prediction, a view reinforced by evolutionary arguments (Kandori, Mailath and Rob, 1993; Young, 1993) and by experimental evidence (Goeree and Holt, 2001), all of which suggest that behaviour converges to the safer outcome even when superior equilibria exist.<sup>1</sup>

Formally, let each player  $i \in \{1, 2\}$  have actions  $A_i = \{\text{Stag}, \text{Hare}\}$ . The payoff bimatrix is written as

	Stag	Hare
Stag	$(u_{11}, v_{11})$	$(u_{12}, v_{12})$
Hare	$(u_{21}, v_{21})$	$(u_{22}, v_{22})$

where  $u_{jk}$  (resp.  $v_{jk}$ ) denotes player 1's (resp. player 2's) payoff when row plays  $j$  and column plays  $k$ . The game is a Stag Hunt when the payoffs satisfy

$$u_{11} > u_{22} > u_{21} \geq u_{12}, \quad v_{11} > v_{22} > v_{12} \geq v_{21}, \quad (1)$$

so that both (Stag, Stag) and (Hare, Hare) are strict pure Nash equilibria while unilateral deviation from Stag is especially costly. From these inequalities it follows by inspection that each player's best reply to Stag is Stag and each player's best reply to Hare is Hare, hence the two diagonals are Nash.

<sup>1</sup>See Harsanyi and Selten (1988); Kandori et al. (1993); Young (1993); Goeree and Holt (2001).

The diagonal (Stag, Stag) is *payoff-dominant* if it Pareto-dominates the other diagonal, that is

$$u_{11} + v_{11} > u_{22} + v_{22}. \quad (2)$$

Risk-dominance is tested by the Harsanyi–Selten risk-product. For the two diagonals define

$$\Delta_{11} = (u_{11} - u_{21})(v_{11} - v_{12}), \quad \Delta_{22} = (u_{22} - u_{12})(v_{22} - v_{21}). \quad (3)$$

Diagonal (Stag, Stag) is risk-dominant if  $\Delta_{11} > \Delta_{22}$ , while (Hare, Hare) is risk-dominant when the inequality is reversed.<sup>2</sup>

This illustrates the tension that proposes that the efficient equilibrium is fragile to unilateral deviation while the safe equilibrium is robust to uncertainty about the opponent.

Recent computational work provides further support for this view. Con-dorelli and Furlan (2025) show that neural networks trained adversarially by regret minimisation across a wide distribution of games consistently select the risk-dominant Nash equilibrium. Their findings raise the question of whether this selection is an intrinsic property of decentralised learning or whether it can be modified through prior exposure. In particular, can a training environment that systematically rewards the payoff-dominant outcome instil a bias strong enough to persist when agents later encounter more general strategic settings?

This chapter addresses that question. It introduces a structured pre-training curriculum in which networks are first exposed exclusively to Stag Hunt games

---

<sup>2</sup>See Harsanyi and Selten (1988) for the linear tracing interpretation and the risk-product test.

with unambiguous payoff-dominant equilibria, before being trained in the standard adversarial regime on randomly generated games. The analysis tests whether the induced preference for efficiency endures, or whether the inherent logic of regret minimisation ultimately reasserts the bias toward risk dominance. The results contribute to the debate on equilibrium selection by clarifying the extent to which learned priors can shape play within decentralised learning dynamics.

## 2.1 Literature Review

There exist many Nash equilibria differentiated by refinements such as payoff-dominance (preferring the Pareto-efficient equilibrium) and risk-dominance (preferring the equilibrium with the largest basin of attraction). Harsanyi and Selten’s equilibrium-selection theory (1988) originally emphasised payoff-dominance, but later work by Kandori, Mailath, and Rob (1993) and Peyton Young (1993) showed that in evolving or learning populations (with small “mutations” or experimentation), risk-dominant equilibria often prevail in the long run. Recent empirical studies with neural agents echo this, such as the Condorelli and Furlan (2025) Paper (Deep Learning Across Games) in which two neural-network agents are trained on a stream of random 2-player games using gradient descent on instantaneous regret, and find the play converges to Nash equilibria in almost all games, tending to select the risk-dominant equilibrium in roughly 80% of coordination games. This large-scale “learning across games” perspective formalises earlier informal arguments (Fudenberg & Levine, 1998; Kreps, 1990) that agents generalise across different but related games. By fitting neural players to minimise regret on a billion random games, Condorelli & Furlan demonstrate that Nash play emerges without repeated play of the same game as their agents’ strategies become  $\varepsilon$  - Nash across new test games (with maximal regret almost

0 for most games).

A large amount of literature in learning theory shows that simple “no-regret” dynamics converge to equilibrium concepts. In repeated play of a fixed normal-form game, if each player applies a regret-minimisation algorithm, the time-average of play converges to the set of coarse correlated equilibria. Crucially, in two-player zero-sum games this guarantee sharpens so that any no-regret dynamics drive the time-average to a Nash equilibrium. For example, Blackwell’s approachability theorem underlies regret-matching dynamics, and Hannan’s seminal result showed that regret-minimising “forecasters” achieve coarse correlated equilibrium. In self-play settings where all agents use such learners, empirical convergence is often fast and smooth. Hybrid algorithms like Fictitious Self-Play and CFR combine regret minimisation with best-response updates. Heinrich and Silver (2016) introduce Neural Fictitious Self-Play (NFSP), an end-to-end deep-RL algorithm that learns approximate Nash equilibria in large imperfect-information games by combining fictitious-play averaging with neural policy learning. NFSP approaches a Nash equilibrium in poker domains where naive RL diverges.

Motivated by real-world applications such as trading and subgame solving where agents face many related games, recent work explores training regret learners on distributions of games. Sychrovský et al. (2024) extend the “learning to optimise” paradigm to regret minimisation where they meta-train neural regret-minimisers offline so they adapt faster in self-play across similar games. Marris et al. (2023) take a different angle in which they train a special equivariant neural network, a Neural Equilibrium Solver, that takes the entire payoff matrix of any 2-player game and outputs an approximate equilibrium (NE, CE

or CCE) in one forward pass. Their model is trained on random games and shows striking zero-shot generalisation to larger unseen games. Similarly, Condorelli and Furlan’s result shows that generic neural players can effectively “learn Nash” across games merely by minimising regret. Therefore, both theory and new ML experiments demonstrate that no-regret learning over heterogeneous games yield Nash play, often favouring the risk-dominant equilibrium in coordination games.

### 3 Methodology: Training Game-Playing Neural Networks

#### 3.1 Neural Networks Architecture

A neural network player is modelled as a differentiable function that maps a bimatrix game into a mixed strategy. Consider two players, row (player 1) and column (player 2). A bimatrix game is a pair  $G = (U^{(1)}, U^{(2)})$  where  $U^{(i)} \in R^{2 \times 2}$  denotes player  $i$ ’s payoff matrix and the entry  $(U^{(i)})_{jk}$  is the payoff to player  $i$  when the row player chooses action  $j \in \{1, 2\}$  and the column player chooses action  $k \in \{1, 2\}$ . A mixed strategy for player  $i$  is a probability vector  $\sigma_i \in \Delta(\{1, 2\})$ .

For a profile  $\sigma = (\sigma_1, \sigma_2)$  the expected payoff to player  $i$  is

$$\pi_i(G, \sigma) = \sum_{a_1, a_2} \sigma_1(a_1) \sigma_2(a_2) U_{a_1 a_2}^{(i)}. \tag{4}$$

The (player-wise) regret of player  $i$  at profile  $\sigma$  in game  $G$  is defined by

$$R_i(G, \sigma) = \max_{a_i \in \{1, 2\}} \pi_i(G, (a_i, \sigma_{-i})) - \pi_i(G, \sigma) \geq 0, \tag{5}$$

and  $\sigma$  is a Nash equilibrium if and only if  $R_i(G, \sigma) = 0$  for both players.

A game-playing neural network for player  $i$  is a continuous, almost-everywhere differentiable map

$$f^{(i)}(\cdot; w^{(i)}) : \mathbb{R}^8 \rightarrow \Delta(\{1, 2\}),$$

where the eight-dimensional input equals the flattened concatenation  $\text{vec}(U^{(1)}), \text{vec}(U^{(2)})$ , and  $w^{(i)}$  denotes the network parameters. The network outputs logits which are transformed by a softmax layer to a valid mixed strategy  $\sigma_i = f^{(i)}(G; w^{(i)})$ .

The architecture used throughout the experiments is a fully connected multilayer perceptron. The standard specification for reported runs comprises two hidden layers of 256 units each with ReLU activation and an output layer of two logits. This functional form is sufficiently expressive to approximate the map from games to strategies while remaining computationally efficient for large batched training.

### 3.1.1 Training Architecture

Training proceeds in an adversarial, across-games regime. Two independent networks, parameterised by  $w^{(1)}$  and  $w^{(2)}$ , are initialised randomly and jointly updated across a sequence of training steps. At each iteration a batch  $\mathcal{B}$  of games is sampled and, for every  $G \in \mathcal{B}$ , the networks produce mixed strategies  $\sigma_i = f^{(i)}(G; w^{(i)})$ . The instantaneous loss for player  $i$  on game  $G$  is the squared personal regret

$$\mathcal{L}_i(G; w^{(i)}, w^{-i}) = (R_i(G, (f^{(i)}(G; w^{(i)}), f^{-i}(G; w^{-i}))))^2. \quad (6)$$

Squared regret is used because zero personal regret characterises Nash equilibria and because the squared form yields numerically stable gradients whose magnitude adapts with the regret.

Given a batch  $\mathcal{B}$ , the network for player  $i$  minimises the average loss

$$\frac{1}{|\mathcal{B}|} \sum_{G \in \mathcal{B}} \mathcal{L}_i(G; w^{(i)}, w^{-i}) \quad (7)$$

With optimiser  $\mathcal{O}$  and learning rate  $\eta_t$ , the parameter update at step  $t$  is the standard first-order update produced by  $\mathcal{O}$  (Adam is used in the reported runs). Denoting by  $w_t^{(i)}$  the parameters at step  $t$ ,

$$w_{t+1}^{(i)} = \text{OptStep}\left(w_t^{(i)}, \nabla_{w^{(i)}} \frac{1}{|\mathcal{B}|} \sum_{G \in \mathcal{B}} \mathcal{L}_i(G; w_t^{(i)}, w_t^{-i}), \eta_t\right). \quad (8)$$

Both networks are updated in parallel; each network’s loss depends on the opponent’s current mapping, so the dynamics are coupled through the dependence of the loss on  $w_t^{-i}$ .

Games are sampled independently from a distribution over bimatrix payoffs. The baseline training distribution is the uniform distribution over a normalised payoff space (zero mean and bounded Frobenius norm) so that sampling does not expose the networks to trivial affine transformations of previous examples.

### 3.1.2 Pre-Training Curriculum

The Stag Hunt pre-training curriculum uses a specialised sampler that generates parametrised Stag Hunt matrices where both pure diagonals are best responses (the conflicting case).

Let  $\mathcal{G}$  denote the space of games of interest for Chapter 1,  $\mathcal{G} = R^{2 \times 2} \times R^{2 \times 2}$ . A *curriculum* is a finite ordered sequence

$$\mathcal{C} = (\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K) \tag{9}$$

where each  $\mathcal{D}_k$  is a probability distribution over  $\mathcal{G}$ . Associated to the curriculum is a schedule of exposure lengths

$$\mathbf{n} = (n_1, n_2, \dots, n_K), \tag{10}$$

where  $n_k \in N$  is the number of gradient-update steps, or equivalently the number of minibatches, devoted to stage  $k$ . Training under curriculum  $\mathcal{C}$  proceeds by sequential empirical risk minimisation. Writing  $w$  for the networks' parameters and  $\mathcal{L}(G; w)$  for the instantaneous loss (squared personal regret), the staged optimisation performs for  $k = 1, \dots, K$

$$w \leftarrow \text{Optim} \left( w, \nabla_w E_{G \sim \mathcal{D}_k} [\mathcal{L}(G; w)], n_k \right), \tag{11}$$

that is,  $n_k$  steps of the chosen first-order optimiser (Adam) using samples from  $\mathcal{D}_k$ .

Let  $\mathcal{D}_{\text{target}}$  denote the evaluation distribution of interest. A curriculum  $\mathcal{C}$  is said to achieve generalisation to  $\mathcal{D}_{\text{target}}$  if after staged training the learned parameters  $w^*$  satisfy a performance bound on  $\mathcal{D}_{\text{target}}$ :

$$E_{G \sim \mathcal{D}_{\text{target}}} [\mathcal{L}(G; w^*)] \leq \varepsilon, \tag{12}$$

for a suitably small  $\varepsilon > 0$ . Generalisation is evaluated by task-relevant statis-

tics such as Mean MaxReg and equilibrium-selection frequencies on held-out test sets sampled from  $\mathcal{D}_{\text{target}}$ .

For the experiments reported in this chapter, a natural instantiation is

$$\begin{aligned} \mathcal{D}_1 &= \text{biased Stag Hunt family with large PD gap,} \\ \mathcal{D}_2 &= \text{parametric Stag Hunt family with smooth variation,} \end{aligned} \tag{13}$$

Pre-training is organised in two stages: an initial stage of exposure to simple biased Stag Hunts with an unambiguous payoff-dominant diagonal, followed by a stage in which the stag/hare/snare parameters are varied smoothly to force generalisation of payoff-seeking behaviour. Main Training occurs on random bimatrix games. Appendix A contains the precise parametrisation used in each stage.

In Stage 1, the payoff-dominant diagonal entries are sampled uniformly from  $[6, 10]$ , while the subordinate diagonal entries are sampled from  $[1, 3]$ . Off-diagonal payoffs are sampled from  $[-5, 0]$ . In Stage 2, the game is a parametric Stag Hunt with a fixed Stag payoff of 5.0 and a Snare of -2.0. The Hare payoff,  $h$ , is increased linearly across the training steps according to

$$h(t) = 1.0 + 3.5 \cdot (t/T_{\text{stage2}}) \tag{14}$$

where  $t$  is the current step within the stage and  $T_{\text{stage2}}$  is the total number of steps in Stage 2.

All training is performed in mini-batches to stabilise gradient estimates. Reported runs use batch size  $B = 256$  for both pre-training and main training. Typical computational budgets for the experiments reported in Chapter 1 are 8,000 pre-training batches (2 million Stag Hunt games) followed by 12,000 main-training batches (3 million random bimatrix games). An exponential learning-rate decay is applied to the base learning rate (Adam with base  $\eta_0 = 10^{-4}$ ).

### 3.1.3 Performance Evaluation

Performance and equilibrium selection are evaluated with the maximum normalised regret statistic. For a given game  $G$  and profile  $\sigma$  the maximum normalised regret is

$$\text{MaxReg}(G, \sigma) = \frac{\max_i R_i(G, \sigma)}{\max_{a_1, a_2} U_{a_1 a_2}^{(i)} - \min_{a_1, a_2} U_{a_1 a_2}^{(i)}}, \quad (15)$$

which rescales regret by the game-specific payoff range. A profile is treated as numerically Nash when MaxReg falls below a small  $\varepsilon$  threshold consistent with observed learning curves. For the purposes of classification, a profile is considered numerically Nash if its *MaxReg* falls below a threshold of  $\varepsilon = 0.01$ .

Equilibrium selection in conflicting Stag Hunt instances is classified operationally by the agents' probability mass on diagonal actions. Denote the payoff-dominant diagonal action by PD and the risk-dominant diagonal action by RD (identified via diagonal utilitarian comparison and the Harsanyi–Selten risk-product criterion respectively). A game is counted as a PD (RD) selection whenever both networks place probability greater than 0.95 on the PD (RD) diagonal; otherwise the game is classified as other/mixed. These thresholds are chosen to match the sharp empirical selection observed in the large-scale

batched evaluation.

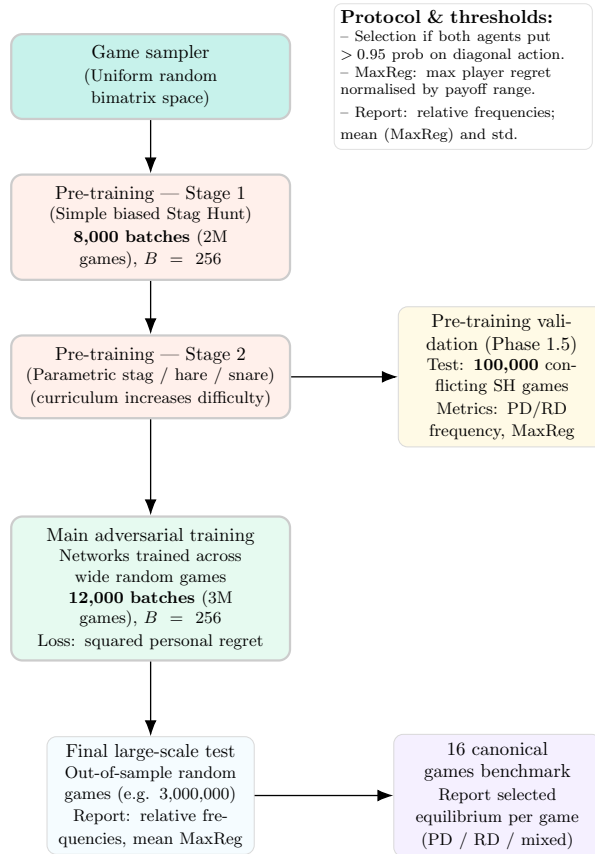


Figure 1: Flow Chart of the curriculum and evaluation pipeline.

## 4 Results

This section reports the empirical findings from the curriculum experiments. Optimisation and convergence diagnostics are reported first, then the effect of the staged pre-training curriculum on conflicting Stag Hunt instances and how this effect evolves under adversarial across-games training, and lastly out-of-sample evaluation on a large random test set and on the sixteen canonical benchmark games.

### 4.1 Optimisation and Convergence Diagnostics

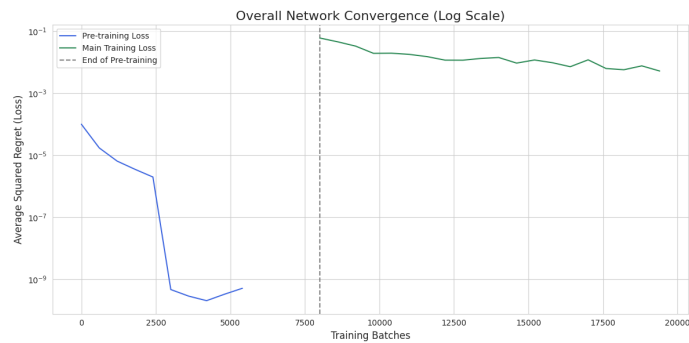


Figure 2: Learning

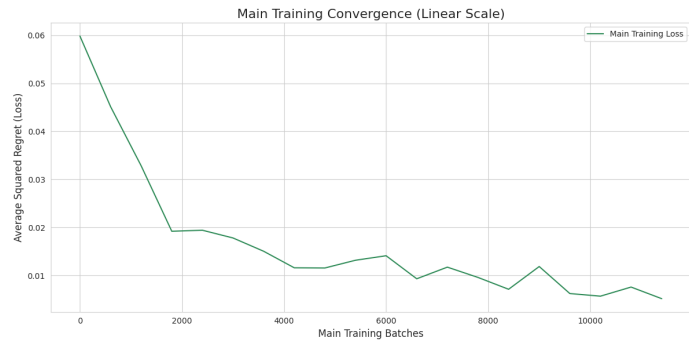


Figure 3: Main Loss

Training is stable and the optimisation converges. The squared-regret loss

decreases monotonically in both pre-training and main-training phases and the learned mappings attain low regret on held-out random games. The aggregate performance on the held-out random sample is summarised in Table 1 below. The reported mean MaxReg is 0.140511 and the standard deviation is 0.208435 thus indicating that its playing approximate Nash behaviour on the bulk of the distribution, while a nontrivial tail of harder instances remains.

Statistic	Value
Mean MaxReg (random test)	0.140511
Std Dev MaxReg (random test)	0.208435

Table 1: Final performance on held-out random games

## 4.2 Effect of the Pre-Training Curriculum

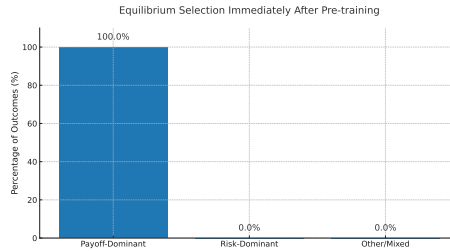


Figure 4: Effect of Pre Training

The staged pre-training curriculum produces a marked bias toward payoff-dominant coordination on the Stag Hunt validation set as seen in the figure above. Immediately after Stage 1 the networks place probability mass on the payoff-dominant diagonal in 93.4% of conflicting instances. After the full Stage 2 curriculum the PD frequency increases to 99.5%. Recorded MaxReg values remain small during pre-training, confirming that induced PD behaviour is compatible with near-equilibrium play. Table 2 reproduces the selection frequencies recorded in the evaluation.

Phase	PD Freq (%)	RD Freq (%)	Other / Mixed (%)
After Stage 1 (simple biased SH)	93.4	0.0	6.6
After full pre-training (Stage 2)	99.5	0.0	0.5

Table 2: Pre-training equilibrium selection on conflicting Stag Hunt validation sets

### 4.3 Transition Under Adversarial Main Training

The results from the evaluation shows that the pre-training induced bias is rapidly eroded under adversarial across-games training. In the logged checkpoints for the reported run, the PD fraction falls to 0.0 within the early main-training steps and remains negligible thereafter. What replaces payoff-dominance varies by instance. In some conflicting games the networks converge to the Harsanyi-Selten risk-dominant diagonal. In other instances, the networks converge to mixed strategies. The recorded traces therefore evidence a qualitative shift away from payoff-dominance rather than a uniform restoration of a single alternative refinement.

### 4.4 Final Performance on Random Games and Canonical Benchmarks

The networks’ final performance is evaluated on a large held-out sample of random  $2 \times 2$  games and on the sixteen canonical games.<sup>3</sup> Table 3 reproduces the canonical-game convergences and the PD/RD classification extracted from the output log.

Where a game prescribes a unique mixed Nash equilibrium, the networks converge to mixed play as theory predicts. Matching Pennies is the clearest example. At the same time several canonical coordination games still show payoff-dominant selection in the reported run. Concretely, the network plays PD in SH1, BoS1 and Chicken2. Two canonical cases, SH3 and BoS2, are

<sup>3</sup>see Condorelli and Furlan 2025

Game	Network convergence	Classification	Picked PD over RD?
SH1	(Top, Left)	Payoff-Dominant (PD)	Yes
SH2	Mixed	Mixed / Other	No
SH3	(Top, Left)	PD and RD coincide (PD=RD)	Both
SH4	Mixed	Mixed / Other	No
MP1	Mixed	Mixed / Other	No
MP2	Mixed	Mixed / Other	No
BoS1	(Top, Left)	Payoff-Dominant (PD)	Yes
BoS2	(Top, Left)	PD and RD coincide (PD=RD)	Both
PD1	(Bottom, Left)	Other (off-diagonal)	No
PD2	Mixed	Mixed / Other	No
Chicken1	(Bottom, Left)	Other (off-diagonal)	No
Chicken2	(Top, Left)	Payoff-Dominant (PD)	Yes
ZS1	Mixed	Mixed / Other	No
ZS2	Mixed	Mixed / Other	No
C1	Mixed	Mixed / Other	No
C2	Mixed	Mixed / Other	No

Table 3: Canonical games: network convergence, classification and whether PD was chosen over RD

such that the payoff-dominant and risk-dominant diagonals coincide and the network chooses those diagonals. Many other canonical games instead end in mixed or off-diagonal outcomes. Thus, while the initial instilled bias erodes in the adversarial stress-test, traces of that bias can be observed in the 16 games where given the choice, the networks chose PD over RD. This also speaks to the broader idea that in unfamiliar environments, such as adversarial training of general games, the networks revert to a combination of mixed or RD outcomes, whereas in more familiar environments to the pre training, it retains traces of the instilled bias and chooses PD over RD.

## 4.5 Summary

The experiments show that the staged curriculum induces a strong payoff-seeking prior on structured Stag Hunt instances. This prior is visible in pre-training validation and in several canonical coordination games. Under adversarial across-games training the prior is rapidly erased in the aggregate random

environment. Long-run outcomes are mixed, or rarely RD depending on local game geometry. However in the benchmark 16 canonical games tested after the out of sample test, the networks prefer PD over RD suggesting that in more familiar environments to the pre training, it retains traces of the instilled bias.

## 5 Chapter Two : Generalising CE

The second chapter extends the analysis beyond Nash equilibria by considering environments where coordination requires conditional play. The objective is to examine whether decentralised neural agents trained by regret minimisation can learn to play and generalise correlated equilibria. While Nash equilibria capture stability under unconditional deviations, correlated equilibria enlarge the attainable set by allowing agents to condition on signals, thereby supporting outcomes that are otherwise inaccessible. The chapter therefore investigates whether modifying objectives and information structures enables decentralised learning to realise these richer patterns of coordination.

A correlated equilibrium in a finite game is a distribution  $\mu \in \Delta(A_1 \times A_2)$  over joint action profiles such that for every player  $i$  and every pair of actions  $a_i, a'_i \in A_i$ ,

$$\sum_{a_{-i} \in A_{-i}} \mu(a_i, a_{-i}) U^{(i)}(a_i, a_{-i}) \geq \sum_{a_{-i} \in A_{-i}} \mu(a_i, a_{-i}) U^{(i)}(a'_i, a_{-i}), \quad (16)$$

meaning that no player can increase their expected payoff by unilaterally deviating from the recommendation they receive. Unlike Nash equilibria, which require independence across players' mixed strategies, correlated equilibria allow coordination via an external signal, enabling outcomes otherwise unattainable under independent play. The theoretical appeal of this concept lies in its ability to capture richer patterns of behaviour, but it is not obvious whether decentralised learners can reproduce such coordination when trained only through experience.

This required a systematic investigation into the conditions under which such learning is possible. The methodology was therefore designed as an iterative pro-

cess of inquiry, beginning with a direct application of existing frameworks and progressively refining the experimental design in response to empirical results. This section details this journey, starting with the exploratory analysis that revealed the limitations of standard regret-minimisation techniques, and culminating in the final, a successful methodology.

The sample space for Chapter 2 consists of pairs  $(G, P)$ , where  $G = (A_1, A_2, u_1, u_2)$  is a finite two-player game and  $P \in \Delta(S_1 \times S_2)$  is a signal distribution over private recommendations. A training example is a draw  $(G, P)$ , followed by a recommendation profile  $s \sim P$ . The learning task is then to map inputs  $(G, s_i)$  into mixed actions in  $\Delta(A_i)$ .

## 5.1 Exploratory Analysis: The Failure of Standard Regret as a Learning Objective

This chapter pursues an iterative programme of inquiry. The investigation begins with a direct application of the regret based learning framework used in Chapter 1 and progressively refines the experimental design in response to empirical failure. The exploratory phase demonstrates that the personal regret objective that generates Nash play repels signal contingent strategies. The negative result motivates a fundamental change in the learning target. The final methodology replaces the personal regret loss with an obedience or swap regret loss and implements a learning environment designed to test for true generalisation of correlated play. What follows describes this trajectory and then gives the precise specification of the final procedure.

## 5.2 Exploratory Phase: Baseline Protocol and Interventions

The exploratory phase retained the learning architecture of Chapter 1 and augmented each network with an input encoding a private recommendation. At each training step a coordination game  $G$  was sampled, a mediator produced a recommendation profile  $s = (s_1, s_2)$  according to a signal distribution  $P$ , and each network received as input, the flattened payoff matrices together with a one-hot encoding of its private recommendation. Training otherwise followed the adversarial protocol of Chapter 1.

Formally, a *coordination game* is a finite two-player game

$$G = (A_1, A_2, u_1, u_2), \tag{17}$$

where  $A_i$  denotes the finite action set of player  $i$  and  $u_i : A_1 \times A_2 \rightarrow R$  is the payoff function of player  $i$ . A pure-strategy profile  $(a_1, a_2) \in A_1 \times A_2$  is a *coordination outcome* if it is a Nash equilibrium and both players receive at least as high a payoff as in any deviation to a non-equilibrium action.

A *signal structure* is defined by a finite set of recommendations  $S_i$  for each player and a joint distribution

$$P \in \Delta(S_1 \times S_2) \tag{18}$$

over recommendation profiles. At each play of the game, a mediator samples  $s = (s_1, s_2) \sim P$  and privately communicates component  $s_i$  to player  $i$ .<sup>4</sup> A

---

<sup>4</sup>The formulation follows Aumann (1974), who introduced correlated equilibrium via a mediator recommending signals to players.

*signal-conditioned strategy* for player  $i$  is a map

$$\pi_i : S_i \times \mathcal{G} \rightarrow \Delta(A_i), \quad (19)$$

which assigns a mixed action to each realised recommendation  $s_i$  conditional on the game  $G$ .

In the experiments reported, the signal is processed as an additional one-hot encoded vector concatenated to the flattened payoff matrices. Hence the input to the network at training step  $t$  is

$$x_t^{(i)} = (\text{vec}(U^{(1)}), \text{vec}(U^{(2)}), \text{onehot}(s_i)), \quad (20)$$

where  $\text{vec}(\cdot)$  denotes flattening to a vector and  $\text{onehot}(s_i)$  encodes the private recommendation.

For completeness, recall the personal regret used in the exploratory experiments. Let  $A_i$  denote player  $i$ 's finite action set and let  $\sigma = (\sigma_1, \sigma_2)$  denote the pair of mixed strategies. The personal regret of player  $i$  in game  $G$  at profile  $\sigma$  is

$$R_i(G, \sigma) = \max_{a_i \in A_i} u_i(a_i, \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}) \geq 0. \quad (21)$$

The loss used in the exploratory phase is the squared personal regret

$$L_{\text{personal}} = (R_i(G, \sigma))^2. \quad (22)$$

All exploratory interventions kept  $L_{\text{personal}}$  as the optimisation target unless stated otherwise. The objective is differentiable in the network parameters under the usual mixed strategy observability assumption and gradients were

computed by backpropagation.

### 5.2.1 Interventions Examined

A broad range of representational, architectural and optimisation interventions were tested. The purpose of these trials was twofold. First, to establish whether the signal was ignored for lack of representation and second, to test whether simple curriculum or initialisation choices could produce agents that continue to obey a mediator under subsequent adversarial training. The interventions are summarised below.

**Signal prominence** The dimensionality of the recommendation embedding was increased from 1 or 2 dimensions to 24 and 64. A learned lookup mapped recommendations to embeddings that were concatenated with the game representation. The input scales were also rebalanced so that the signal embedding dominated the input norm. These changes increased mutual information between the signal and early layer activations. They did not, however, produce signal contingent play under  $L_{\text{personal}}$ . Networks learned representations that neutralised the signal before the final layer and output logits remained effectively independent of  $s$ . Table 4 below provides summary statistics, and further graphs from this experiment can be found in Appendix B.

Metric	Value
Final Success Rate	0%
Final Avg. Policy Divergence	0.0000
Final Avg. $P(\text{Coop} \mid S = 0)$	1.0000
Final Avg. $P(\text{Coop} \mid S = 1)$	1.0000

Table 4: Final Summary Statistics - Signal

**Multiplicative interaction and gating** Interactions between game and signal embeddings were forced. In one experimental variant, the game and signal embeddings were first mapped to vectors of equal dimension and then combined using elementwise multiplication. In a second variant, a learned gate  $g(s, G) \in [0, 1]^d$  multiplicatively reweighted the game embedding. Both designs made it impossible for the network to process payoffs without processing the signal. Despite this, during optimisation with  $L_{\text{personal}}$ , networks learned gating weights or multiplicative coefficients that erased signal information prior to the output layer. The dynamics found low regret directions that ignored the recommendation. Below is a table of summary statistics and further graphs relating to this experiment can be found in Appendix B.

Metric	Value
Final Success Rate	0%
Final Avg. Policy Divergence	0.0000
Final Avg. $P(\text{Coop} \mid S = 0)$	1.0000
Final Avg. $P(\text{Coop} \mid S = 1)$	1.0000

Table 5: Final Summary Statistics - Gating

**Curriculum learning with Pareto valuable signals** Staged training were tested in which the early stage exposed agents only to games where following the mediator produced strictly higher payoffs for both players than any Nash equilibrium. Stage two exposed the networks to the usual wide distribution of games. Both hard filtering and soft schedules were tested in which the probability of Pareto valuable instances decayed linearly. The probability of sampling a Pareto-valuable game,  $p_{\text{pareto}}(t)$ , decayed linearly from 1.0 to 0 over the pre-training phase according to

$$p_{\text{pareto}}(t) = 1 - (t/T_{\text{pre-train}}) \tag{23}$$

where  $t$  is the training step. Agents learned to follow the mediator during the restricted stage. On transition to the full distribution, however, adversarial updates rapidly pushed policies into the nearest Nash basins. The pre-trained obedience did not persist in the vast majority of initialisations. Find below a table of summary statistics and further graphs relating to this experiment can be found in Appendix B.

<b>Metric</b>	<b>Value</b>
Final Success Rate	43%
Final Avg. Policy Divergence	0.5758
Final Avg. $P(\text{Coop} \mid S = 0)$	0.7017
Final Avg. $P(\text{Coop} \mid S = 1)$	0.2946

Table 6: Final Summary Statistics - Curriculum

**Stochastic counterfactuals** The baseline exploratory experiments evaluated counterfactuals using the opponent’s mixed strategy outputs rather than action samples. To test realism, experiments were repeated where  $E[\sigma_{-i} \mid s_i]$  was approximated by 10 Monte Carlo sampling of opponent actions. Learning became slower and more fragile but the qualitative conclusion remained. Personal regret training with sampled counterfactuals continued to repel obedient states. Further graphs pertaining to this experiment can be found in Appendix B.

<b>Metric</b>	<b>Value</b>
Final Success Rate	0%
Final Avg. Policy Divergence	0.0000
Final Avg. $P(\text{Coop} \mid S = 0)$	1.0000
Final Avg. $P(\text{Coop} \mid S = 1)$	1.0000

Table 7: Final Summary Statistics - Stoichastic

### 5.3 Theoretical Foundation and the Obedience Regret Objective

The failure of the exploratory interventions points to a conceptual mismatch. Personal regret measures the incentive for unconditional unilateral deviation and is minimised at Nash. Correlated equilibrium is defined by incentive constraints conditional on signals. The classical link between internal regret and correlated equilibrium motivates the alternative objective. In repeated play, no internal-regret dynamics lead to the correlated equilibrium set while no external-regret dynamics do not.<sup>5</sup>

Formally, the *external regret* of a player over  $T$  periods is defined as

$$R_i^{\text{ext}}(T) = \max_{a'_i \in A_i} \frac{1}{T} \sum_{t=1}^T \left( u_i(a'_i, a_{-i}^t) - u_i(a_i^t, a_{-i}^t) \right), \quad (24)$$

which measures the gain the player could have obtained by committing ex ante to a single alternative action  $a'_i$ .

By contrast, the *internal regret* compares each action  $a_i$  actually played to alternative actions  $a'_i$ , defining for each pair  $(a_i, a'_i)$

$$R_i^{\text{int}}(a_i \rightarrow a'_i, T) = \frac{1}{T} \sum_{t: a_i^t = a_i} \left( u_i(a'_i, a_{-i}^t) - u_i(a_i, a_{-i}^t) \right), \quad (25)$$

and the internal regret is the maximum of these quantities across all pairs. Minimising external regret guarantees convergence to the set of coarse correlated equilibria, while minimising internal regret guarantees convergence to the set of correlated equilibria. In this sense, swap regret emerges as a strengthened form of internal regret that provides a practical learning criterion for inducing

<sup>5</sup>See Hart and Mas-Colell (2000) and Foster and Vohra (1997).

correlated play.

This insight was operationalised by defining an obedience or swap regret that measures, for each recommendation, the maximal positive gain from swapping from the recommended action to any alternative when evaluated against the opponent’s conditional behaviour.

Formally let  $S$  denote the finite set of recommendations. For a given recommendation  $s_i$  define the opponent’s conditional expectation of play as

$$E[\sigma_{-i} | s_i] = \sum_{s_{-i} \in S} P(s_{-i} | s_i) f_{\theta_{-i}}(s_{-i}, G, P), \quad (26)$$

where  $f_{\theta_{-i}}$  denotes the opponent’s meta policy. Let  $a(s_i)$  denote the mediator’s recommended pure action at signal  $s_i$ . The obedience regret at  $s_i$  is

$$R_i^{\text{obed}}(s_i) = \max_{a' \in A_i} \max\{0, u_i(a', E[\sigma_{-i} | s_i]) - u_i(a(s_i), E[\sigma_{-i} | s_i])\}. \quad (27)$$

This quantity is the maximal positive gain from swapping away from the recommended action when evaluated against the opponent’s conditional mixed strategy. The squared obedience regret loss that is minimised in the final protocol is

$$L_{\text{obed}} = \frac{1}{|S|} \sum_{s_i \in S} (R_i^{\text{obed}}(s_i))^2. \quad (28)$$

Minimising  $L_{\text{obed}}$  forces the learner to internalise conditional incentive constraints. The loss is zero only when no agent, for any recommendation, has a profitable unilateral swap. This is the defining condition of a correlated equilibrium.

## 5.4 Final Training Environment and Data Generation

The final experiments train networks on a dynamically generated family of coordination games. At each training step the generator creates a new coordination game  $G$ , together with a welfare maximising signal distribution  $P$ . Games are constructed so that two pure strategy equilibria exist and coordination is non trivial. The signal  $P$  is computed to recommend the welfare maximising profile or a randomised recommendation that implements the welfare improving correlated plan. A welfare-maximising signal  $P$  is a joint distribution over action profiles  $A = A_1 \times A_2$  that places all mass on the profile  $a^*$  maximising utilitarian welfare, defined as the sum of players' payoffs. Formally,

$$a^* = \arg \max_{a \in A_1 \times A_2} (U_a^{(1)} + U_a^{(2)}). \quad (29)$$

The mediator then recommends to each player  $i$  the component  $s_i = a_i^*$ . The signal thus acts as a Pareto-efficient coordination device.

Sampling a fresh pair  $(G, P)$  at each step prevents memorisation. The learning task is to learn a universal mapping  $f_{\theta_i} : (s_i, G, P) \mapsto \sigma_i$  that generalises to novel games and signal distributions. In robustness exercises, the informativeness of  $P$  was varied, introduce noise in the signal, and alter the class of coordination games to verify that successful learning is not an artefact of a specialised generator.

## 5.5 Network Architecture and Optimisation

Each agent is a feed forward neural network that implements the policy  $f_{\theta_i}$ . The input is a concatenation of vectorised game payoffs, a vector representation of the signal distribution  $P$ , and an embedding of the private recommendation  $s_i$ .

The core architecture uses two fully connected hidden layers of 128 units, with rectified linear activations and a softmax output layer that produces a mixed strategy over the two actions. For 2 by 2 coordination games, the output layer has two logits. Ablations test multiplicative interaction variants and deeper and wider networks.

Training proceeds in individual games. For each instance, each network produces its signal conditioned strategy. The opponent’s conditional expectation  $E[\sigma_{-i} | s_i]$  is formed using the opponent network’s outputs and the conditional probabilities  $P(s_{-i} | s_i)$ . The obedience regret  $R_i^{\text{obed}}(s_i)$  is computed for every instance and gradients of the squared obedience regret loss are computed with respect to the network parameters. Parameters are updated with the Adam optimiser. The initial learning rate and decay schedule are chosen to stabilise adversarial updates.

A practical implementation detail is that the evaluation of  $E[\sigma_{-i} | s_i]$  uses the opponent network’s mixed strategy output rather than action realisations. This choice removes sampling noise from the loss and mirrors the mixed strategy observability assumption used in Chapter 1.

## 5.6 Evaluation Protocol and Diagnostics

Generalisation is the primary object of interest. Evaluation is conducted on held out coordination games never seen during training. Two diagnostics measure success.

**Signal Contingency** For each held out game  $G$ , evaluate the learned policy at each recommendation and compute the L2 difference

$$\Delta_i(G) = \|\sigma_i(s = 0) - \sigma_i(s = 1)\|_2. \quad (30)$$

A policy is signal contingent in  $G$  if  $\Delta_i(G) > \varepsilon$  for a small threshold  $\varepsilon$ .

**Correct mapping** A policy maps correctly if the highest probability action under each recommendation coincides with the mediator’s recommended action.

Formally for each signal  $s$  we require  $\arg \max \sigma_i(s) = a(s)$ .

The generalisation success rate is the fraction of held out games for which both diagnostics are satisfied for both players.

Standard learning statistics are also reported. The average of squared obedience regret is plotted against optimisation steps to document convergence. The distribution of per recommendation obedience regret on the test set is reported. MaxReg is measured as in Chapter 1 to show the relationship with Nash stability.

## 5.7 Iteration and justification

The final protocol is the product of the iterative programme described above. The exploratory phase established that minimisation of squared personal regret is not only insufficient for correlated learning in the cross game adversarial training learning environment, but is actively repulsive to obedience. Representational changes, multiplicative interactions, and curriculum schedules altered transients but not the final attractor. These negative results motivate the change in objective. Replacing personal regret with squared obedience regret is a principled realignment of the learning target with the conditional incen-

tive constraints that define correlated equilibrium. The empirical results in the next section show that this change is sufficient to produce signal contingent, generalisable correlated play under adversarial training.

## 6 Final Methodology

### 6.1 The Learning Objective: Obedience (Swap) Regret

The definitive failure of the standard regret hypothesis necessitated a fundamental change in the learning objective. To incentivise coordination on the signal, a loss function based on **obedience regret**, also known as **swap regret** was adopted. This re-frames the agent’s decision problem from the non-cooperative question of ”What is my best overall strategy?” to the coordination-centric question: ”Given my signal, should I obey it?”

The obedience regret for player  $i$  receiving signal  $s_i$  is the incentive to disobey the recommendation  $a_i = s_i$  and ”swap” to an alternative action  $a'_i$ . The loss is the square of the maximum positive regret over all possible swaps. This counterfactual is evaluated based on the agent’s belief about the opponent’s strategy, which is formed by assuming the opponent will follow the policy dictated by its own network for any signal it might receive.

Formally, the loss for player  $i$  given signal  $s_i$  is:

$$\mathcal{L}_{\text{obedience}}(E[\sigma_{-i}]) = \left( \max_{a'_i \in A_i} \max(0, u_i(a'_i, E[\sigma_{-i}]) - u_i(s_i, E[\sigma_{-i}])) \right)^2 \quad (31)$$

where the opponent’s expected strategy  $E[\sigma_{-i}]$  is calculated as:

$$E[\sigma_{-i}] = \sum_{s_{-i} \in A_{-i}} P(s_{-i}|s_i) f_{\theta_{-i}}(s_{-i}, G, P) \quad (32)$$

This loss is only minimised when no player, for any signal it might receive, has a positive incentive to unilaterally deviate from the mediator’s recommendation. This is the precise mathematical definition of a Correlated Equilibrium. This change aligns the agents’ optimisation objective directly with the target equilibrium concept and represents the key theoretical turning point of this research, shifting the goal from individual rationality to coordinated rationality.

## 7 Results

The results demonstrate that minimising squared obedience (swap) regret under the adversarial training protocol produces stable convergence and strong out-of-sample generalisation to previously unseen coordination games. Below reported is convergence behaviour, cross-game generalisation diagnostics and qualitative policy structure.

### 7.1 Convergence of the Obedience Loss

Training proceeded for a total of 900,000 adversarial steps and intermediate diagnostics were logged at intervals of 9,000 steps. The squared obedience regret loss decayed extremely rapidly. At the first logged checkpoint, step 9,000, the recorded loss value is reported as 0.000000 and it remains at, or numerically indistinguishable from, zero at all subsequent logged checkpoints. This rapid collapse to zero indicates that the obedience constraints of no profitable one-shot swap deviations given the agent’s beliefs about opponents, are satisfied early and remain satisfied for the remainder of training. The plotted loss, displayed on a log scale, corroborates a near-monotonic decrease to numerical zero and



Figure 5: Loss

no evidence of systematic oscillatory or divergent dynamics under the Adam optimiser and the chosen learning-rate schedule.

## 7.2 Generalisation success on Held-Out Games

Because generalisation is the principal object of the study, the learned policies on batches of 100 previously unseen games at each checkpoint are evaluated. Success on a held-out game requires two simultaneous conditions which are (i) signal contingency, operationalised as the L2 distance between the player’s conditional strategies exceeding the small threshold  $10^{-3}$ , and (ii) correct mapping, i.e. the highest probability action under each recommendation matches the mediator’s recommended action. Over 100 logged evaluations collected through training, the recorded success rates have the following empirical distribution described in the table below.

Statistic	Value
Number of evaluation checkpoints	100
Mean success rate	79.08%
Median success rate	79%
Population standard deviation	3.71 percentage points
Minimum observed success rate	68%
Maximum observed success rate	90%

Table 8: Summary statistics for success rates across 100 held-out evaluations.

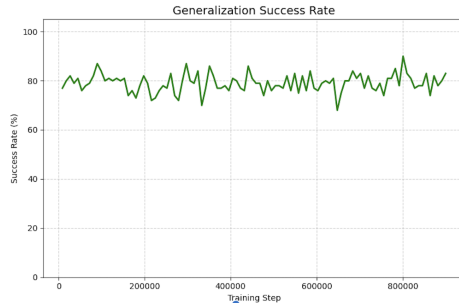


Figure 6: Generalisation Rate

Additional summary counts reinforce the robustness of the outcome. 89/100 logged checkpoints achieved success rates at or above 75%, and 47/100 achieved at or above 80%. Only 2/100 checkpoints fell at or below 70%. These statistics show that high generalisation performance is not a transient phenomenon confined to a few checkpoints. Therefore, the system spends the large majority of checkpoints in a narrow band around 75–90% success.

The success rate on held-out games does not increase monotonically with training steps, but instead oscillates in a relatively narrow band between 75% and 90%. This behaviour is expected given the adversarial training protocol. At each step the generator produces a fresh coordination game with its associated signal distribution, and the evaluation routine likewise samples new games for testing. The generalisation metric therefore reflects both the state of the networks and the idiosyncrasies of the sampled test set. In particular, certain games are intrinsically harder to coordinate on, for example, when payoffs are nearly symmetric across equilibria or when the signal is only weakly informative. When the held-out set at a given checkpoint happens to contain more such “hard” games, the observed success rate is temporarily depressed even though the underlying mapping remains stable. Conversely, when the held-out set contains relatively “easy” games the success rate peaks. The oscillations

thus represent variance across game draws rather than instability of the learning process. This interpretation is corroborated by the fact that the obedience loss converges monotonically to zero and remains there, while the divergence and policy evolution diagnostics, discussed in the next subsection, show stable, persistent signal-contingent behaviour. Hence, from this it can be inferred that the learning dynamics are stable, and the evaluation metric exhibits stochastic fluctuations because it samples across a heterogeneous family of coordination games.

### 7.3 Conditional (signal-contingent) policies

The learned policies are signal-contingent. The L2 divergence metric used in evaluation, the L2 norm between the policy vector for signal 0 and for signal 1, is strictly positive for successful held-out games by construction (the evaluation routine requires  $\Delta > 10^{-3}$  to count a game as signal-contingent).

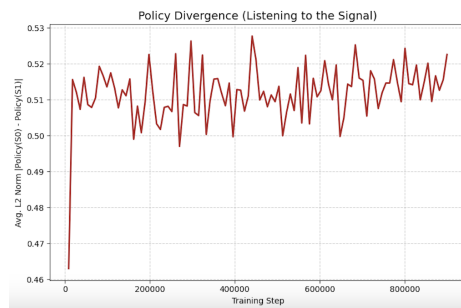


Figure 7: Policy Divergence

The figures above show that, at convergence, conditional probabilities separate cleanly such that the probability of the mediator-recommended action is high when that action is recommended and low otherwise.



Figure 8: Policy Evaluation

The final illustrative bar chart produced for a novel game displays near-deterministic obedience in the sampled instance where the recommended action receives probability close to 0.80, while the non-recommended action receives probability close to 0.2 if the recommended signal is to cooperate and flips to approximately 0.6 and 0.4 when the recommendation is to defect. This could be because the payoff of cooperating is significantly larger for the chosen novel game. This outcome is entirely consistent with the observed convergence of the squared obedience regret loss to numerical zero, and it confirms that the networks implement a robust signal-contingent mapping rather than a signal-agnostic or memorised strategy.

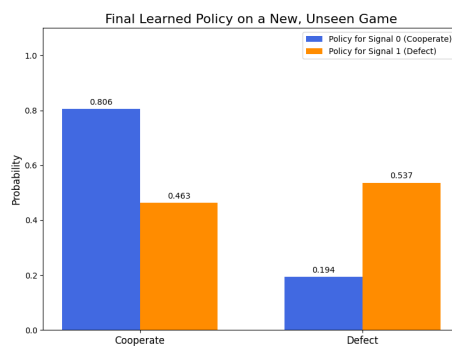


Figure 9: Learned Policy

## 7.4 Summary

The training generator produces a diverse family of  $2 \times 2$  coordination payoff matrices such that each logged training step uses a freshly sampled game and its welfare-maximising signal distribution. The persistently high held-out success rates across 100 logged checkpoints and across the sampled test games indicate that the learned mapping is not a brittle memorisation of a small template set, but rather a universal mapping that transfers across the generator’s support. Importantly, the loss was computed using the opponent’s mixed strategy outputs such that they are not sampled action realisations, which removes sampling variance from the regret estimates and isolates the effect of the obedience objective itself from which it is observed that sampled-action counterfactuals yield qualitatively similar results but with slower learning.

Taken together, these results support the central claim of the chapter which is that replacing a personal-regret objective with a squared obedience (swap) regret objective realigns learning with the conditional incentive constraints that characterise correlated equilibrium and yields stable, generalisable, signal-contingent behaviour under adversarial training. Convergence to numerical zero obedience regret occurs very early by the first logged checkpoint at step 9,000, and held-out success remains robust thereafter with mean success at roughly 79.1%, the minimum at 68%, maximum at 90%, and stdev at 3.71. The networks therefore internalise the mediator’s recommendations and generalise that conditional mapping across a wide family of coordination games.

## 8 Literature Review

Formally, Aumann defined CE as a joint distribution over actions that could arise if a “correlation device” privately recommends actions to players such that each player’s recommendation is a best response given the known recommendation distribution to others. Hart and Mas-Colell (2000) articulate this as, “a CE is the Nash equilibrium of the game augmented with signals, or equivalently a distribution on action profiles induced by a mediator’s instructions”. Crucially, simple learning rules can converge to this broader equilibrium set. Hart & Mas-Colell’s regret-matching procedure uses only each player’s external regrets and adapts probabilities proportionally to regret, and they prove the empirical play converges almost surely to the set of correlated equilibria. More generally, if every player in repeated play minimises swap (internal) regret (i.e. they regret not having consistently replaced each action with each other possible action), then the joint time-average converges to the set of CEs. Blum & Mansour (2007) formalise this: if each player’s swap regret grows sublinearly ( $R/T \rightarrow 0$ ), then the empirical distribution of play is an approximate correlated equilibrium. Thus, no internal-regret learning is both a necessary and sufficient condition for coarse or correlated equilibrium in general games (unlike external regret, which only guarantees coarse correlation).

Theoretical work on implementing CE in games with communication shows that a planner or even cheap talk can achieve any CE. Practically, multi-agent learning algorithms can simulate such correlating devices. For example, Cigler & Faltings (2013) consider a “channel allocation” anti-coordination game where identical agents must choose distinct resources. They show that if agents share a common random signal (even a simple integer from  $1 \dots K$ ), and condition their action on it, learning dynamics converge to an efficient CE. Specifically,

their agents learn separate pure equilibria for each signal value, and the joint outcome (averaged over signals) is a fair, correlated allocation. This highlights a key point which is that even a “dumb” coordination signal can replace a smart mediator if agents learn to interpret it appropriately.

Multi-agent reinforcement learning (MARL) models explore related ideas under communication or signalling. Chen et al. (2022) introduce a coordination signal into cooperative MARL where they append a public random variable, and train agents with a new Signal-Instructed Coordination (SIC) module. The signal encourages agents to develop conditional policies that can coordinate better. Empirically, SIC integration significantly improves performance on matrix games and standard benchmarks, confirming that adding a correlated signal can overcome decentralised limitations. Likewise, Li et al. (2024) propose AgentMixer, a network architecture that nonlinearly combines each agent’s local policy into a joint policy, explicitly allowing coordination akin to a correlating device. They prove AgentMixer’s joint policy converges to an approximate CE, and show it matches or outperforms baselines in benchmarks. These works embody the principle that learning a signal-conditioned policy (or mixing policies via a learned “mixer”) can unlock the gains of correlation without a central coordinator.

Mohri & Yang extend the swap-regret literature in a way that is particularly germane to any empirical programme aiming for signal-contingent coordination. They define conditional swap regret, allowing swap comparisons that are conditioned on bounded action histories, and prove that minimising conditional swap regret converges to a correspondingly stronger solution concept — the conditional correlated equilibrium. Importantly for applied learning, they

provide algorithms for minimising conditional swap regret with bounded conditioning and extend these guarantees to partial-information (bandit) scenarios. Most importantly, they give a principled formalisation that obedience should be assessed not only globally but conditioned on the recommendation or signal history, and doing so has both theoretical guarantees and algorithmic pathways.

Therefore, as seen from the above discussion, the correlated equilibrium and recommendation literatures converge on the idea that signal-contingent learning can implement coordination. Foundational theorems (Aumann 1974, Hart & Mas-Colell 2000, Foster & Vohra 1997) show that no-internal-regret (or calibrated) learning leads to CE. In applied settings, algorithms like Correlated-Q (Greenwald & Hall, 2003) explicitly compute CEs during learning, and newer deep-Reinforcement Learning methods augment agents with communication channels or conditioning signals. Empirical studies confirm that including learned signalling mechanisms significantly broadens achievable equilibria and improves robustness. For instance, agents using coordination signals reliably converge to high-reward equilibria even when independent learning fails. Together, these theoretical and applied results underscore that recommendations or public signals, whether hand-designed or emergent, can guide decentralised learners toward correlated solutions that Pareto-dominate the individualistic Nash outcomes.

## 9 Conclusion

This dissertation investigated how neural-network agents trained by regret-minimisation form and sustain equilibrium play in two complementary settings (i) whether a structured pre-training curriculum can instil and preserve a payoff-dominant bias under subsequent adversarial across-games training; and

(ii) whether decentralised learners can be aligned to learn and generalise correlated equilibria by changing the learning objective from personal regret to obedience (swap) regret.

Chapter One shows that a carefully designed pre-training curriculum effectively instils a strong payoff-seeking prior on structured Stag Hunt instances. Under the staged curriculum, networks place very high probability mass on payoff-dominant diagonals on the pre-training validation set (PD frequency rising to 99.5% after the full curriculum in the reported run), and recorded MaxReg values remain small during pre-training, indicating near-equilibrium behaviour. Nonetheless, when the networks are transferred to the broad adversarial training distribution, this induced bias is rapidly eroded as seen in the reported run where the PD fraction falls to essentially 0% early in main training and remains negligible thereafter, and final equilibrium selections in conflicting games are classified as other/mixed in aggregate. The mean maximum-normalised regret on held-out random games indicates approximate Nash behaviour (reported Mean MaxReg approximately 0.1405, with nontrivial variance), but the aggregate equilibrium selection reverts to mixed or risk-sensitive outcomes once networks experience the broader distribution of strategic environments. These findings demonstrate that pre-training can transiently create an efficiency bias but that adversarial across-games training tends to re-assert the attractors associated with the broader game distribution.

Chapter Two addresses a distinct limitation revealed by the exploratory experiments where minimising squared personal regret actively repels obedience to mediator recommendations and therefore does not produce signal-contingent correlated behaviour under adversarial training. Motivated by the theoretical

link between internal (swap) regret and correlated equilibria, the final protocol replaces the personal-regret objective with a squared obedience (swap) regret loss. Empirically, minimising this obedience loss produces rapid convergence (obedience loss collapses to numerical zero by the first logged checkpoint at step 9,000 in the reported runs) and robust out-of-sample generalisation achieving held-out success rates across 100 logged checkpoints display a mean of approximately 79.1% (median 79%, stdev approximately 3.7 percentage points), with most checkpoints achieving success rates between 75% and 90%. The learned policies are demonstrably signal-contingent and map recommendations to high-probability mediator actions in the majority of test games.

Together, the two chapters highlight two important insights. First, training history and curriculum can create strong inductive priors, but these priors are fragile when the learning environment broadens where the attractors of adversarial across-games dynamics dominate aggregate long-run behaviour. Second, aligning the optimisation target with the object of interest is crucial. Swap/obedience regret is the natural differentiable surrogate for correlated equilibrium, and minimising it yields stable, generalisable signal-contingent policies where personal regret fails.

## References

- [1] Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1), 67–96.
- [2] Blum, A., & Mansour, Y. (2007). From external to internal regret. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT)* (pp. 621–636).
- [3] Chen, Y., Ma, X., Zhang, W., & Yu, Y. (2022). Signal-instructed coordination in multi-agent reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.
- [4] Cigler, L., & Faltings, B. (2013). Decentralized resource allocation in networks using correlated equilibria. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp. 1285–1286).
- [5] Condorelli, D., & Furlan, M. (2025). Deep Learning Across Games. arXiv preprint arXiv:2409.15197.
- [6] Foster, D., & Vohra, R. (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1–2), 40–55.
- [7] Fudenberg, D., & Levine, D. K. (1998). *The Theory of Learning in Games*. MIT Press.
- [8] Goeree, J. K., & Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5), 1402–1422.
- [9] Greenwald, A., & Hall, K. (2003). Correlated-Q learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*.

- [10] Harsanyi, J. C., & Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press.
- [11] Hart, S., & Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5), 1127–1150.
- [12] Heinrich, J., & Silver, D. (2016). Deep reinforcement learning from self-play in imperfect-information games. In *Proceedings of the NIPS Deep Reinforcement Learning Workshop*.
- [13] Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica*, 61(1), 29–56.
- [14] Kreps, D. M. (1990). *Game Theory and Economic Modelling*. Oxford University Press.
- [15] Li, J., Zhao, M., Wang, Y., & Zhang, T. (2024). AgentMixer: Cooperative policy learning via correlated equilibria. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*.
- [16] Ling, C. K., Li, M., Lanctot, M., & Zhang, Y. (2021). Self-play PSRO: Toward optimal multiagent learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.
- [17] Marris, L., McAleer, S., Kroer, C., & Brown, N. (2023). Equivariant neural equilibrium solvers. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*.
- [18] Ortner, R., Salz, T., & Schwarz, M. (2023). Beyond no-regret: Fast convergence to equilibrium through utility maximization. *Journal of Machine Learning Research*, 24(215), 1–34.
- [19] Sychrovský, T., Moravčík, M., Pevný, T., & Lisý, V. (2024). Meta-training regret minimizers across games. *Artificial Intelligence*, forthcoming.

- [20] Tsai, J., & Han, Y. (2020). Optimal reliable strategies in general-sum games. *Theoretical Computer Science*, 812, 57–70.
- [21] Young, H. P. (1993). The evolution of conventions. *Econometrica*, 61(1), 57–84.