A Work Project, presented as part of the requirements for the Award of a Master's degree in Finance and Financial Markets - Executive from the Nova School of Business and Economics.

Learning by Teaching: Reinforcement Learning and Regime-Aware Option Strategy Selection

David António Poceiro do Carmo

Work project carried out under the supervision of:

Gonçalo Sommer Ribeiro

.

1. Abstract

This project applies Reinforcement Learning to options trading, framing the environment as a systematic trading decision process. A Proximal Policy Optimization agent selects among risk-defined option strategies: Vertical Spreads, Straddles, Strangles, Condors, and No-Trade, based on market features such as Greeks, implied volatility, moneyness, and macro indicators (VIX3M/VIX, yield-curve slope, SPX moving averages). Trained on SPY weekly options with a constant five-week Friday-to-Friday maturity from 2020 to 2024, the model captures about 88% of the oracle benchmark and converges to two behavioural modes: short-volatility exploitation and defensive long-volatility positioning, demonstrating it can internalize financially coherent, regime-aware trading logic.

Keywords: Options Trading, Reinforcement Learning, Proximal Policy Optimization, Regime Awareness, Quantitative Strategies

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

2. Introduction

This work explores how Reinforcement Learning (RL) can be applied to options trading as a potential path to alpha, focusing on understanding when and why strategies work rather than on prediction. Consistent profitability depends on recognizing structural market contexts, or "regimes". RL, by optimizing reward through interaction, provides a natural framework for this process because trading decisions unfold sequentially and rewards are realized over time. Within this framework, Proximal Policy Optimization (PPO) is chosen for its stable learning process and ability to adapt consistently to changing conditions. The agent is trained to select among standard option strategies and examined in order to study whether a coherent, profitable trading logic can emerge autonomously. Options trading offers an ideal environment for this investigation since each strategy expresses a distinct view on market direction and volatility and structured payoffs and risk–return profiles that can be clearly defined. In this controlled setting learning outcomes are measurable and interpretable, allowing the study of how the agent recognises market regimes and internalises when different strategies succeed.

This report captures the current checkpoint in the broader ambition to develop an intelligent learning system capable of selecting the strategy with the highest probability of success based on its past trading experience. It also represents a learning checkpoint for us personally, as we deepened our own understanding of the topic by designing, framing, and observing how the agent learns.

The mathematical foundations of PPO are well established in the reinforcement-learning literature (Schulman et al. 2017; Sutton and Barto 2018), and the algorithm has since become a standard baseline widely applied across diverse domains. As the focus of this project lies in the financial interpretation of learning behaviour rather than in algorithmic development, the discussion is limited to the framing of the financial environment where the agent will learn.

3. Market regimes

Financial markets exhibit distinct patterns of behaviour shaped by volatility, directionality, and risk appetite. Market regimes reflect these recurring configurations and are central to understanding the performance of options trading strategies because their payoff structures are designed to exploit specific combinations of trend and volatility (Chujoy, Seth, and Chantarajirawong 2016). Directional markets, where prices move persistently upward or downward, tend to favour strategies with defined directional exposure, such as Vertical Spreads. Periods of volatility expansion or contraction, alter the relative attractiveness of long and short-volatility positions, shaping the performance of Straddles, Strangles, and Condors (Liu, Li, and Fan 2025). During low volatility markets, credit structures such as Iron Condors offer attractive risk-adjusted returns by collecting premium in exchange for limited tail risk, whereas in volatile or corrective phases, debit strategies become more appealing as their asymmetric payoffs provide greater upside potential relative to the premium paid.

Understanding regimes is not merely descriptive but essential to recognize when a strategy's payoff logic remains valid and to inform decision-making. This anchors the central idea of learning as the process of internalizing the market conditions under which specific strategies succeed or fail, rather than forecasting prices.

In this context, the agent's performance must be analysed both in aggregate and by market regime, using classifications of volatility and trend conditions to assess whether the learned policy exhibits regime-dependent behaviour.

4. Reinforcement Learning

RL is a branch of machine learning that leverages interaction with the environment to make sequential decisions that maximize long-term rewards (Sutton and Barto 2018). Unlike supervised learning, which relies on labelled examples, or unsupervised learning, which

searches for patterns in static data, RL improves performance through experience, opening the door for continuous improvement and adaptation to a changing environment (the ultimate ambition of this project). This feedback loop creates an adaptive decision process in which exploration of new possibilities is balanced with the exploitation of accumulated knowledge (Sutton and Barto 2018, 26–28). The PPO policy gradient design aligns with the objective of learning stable decision behaviour rather than forecasting specific outcomes, as in value-based approaches. While value-based algorithms attempt to estimate the future value of each possible action, PPO follows an actor–critic architecture in which the policy network learns the decision rule directly and the value network provides an adaptive baseline for evaluating outcomes. This structure improves learning stability by restricting the size of each policy update to prevent overreaction to noise or rare events (Schulman et al. 2017), a valuable property when dealing with instruments with the potential for extreme returns, such as options, and in markets susceptible to regime shifts (Terven 2025). As such, PPO provides a controlled mechanism for gradual adaptation, enabling the agent to learn consistent decision rules from volatile and non-stationary data.

The PPO framework implemented in this project combines three structural elements and three functional components designed to mirror trading logic. The environment simulates historical market conditions; the agent represents the decision policy that selects among strategies to maximize expected reward; and the oracle provides realized outcomes such as price movements and option payoffs as feedback. The state representation integrates macro, trend, and directional features; the action space comprises predefined option strategies, allowing the agent to identify which best fits which conditions; and the reward function (corresponding to the realized profit or loss of each strategy held to expiry) serves as the agent's sole learning signal. Together, these components frame the agent's interaction with market data, allowing trading logic to emerge directly from that experience.

5. Environment Design

The RL environment is designed to replicate the decision process of a systematic options trader. Each trading day is treated as a self-contained episode in which the agent evaluates predefined option strategies candidates and decides whether and how to trade. This structure mirrors the repeated evaluation of opportunities that characterize options trading. To reduce noise and ensure comparability, the action space is limited to standard option strategy templates, focusing learning on strategy selection rather than configuration or sizing. It includes Vertical Calls, Vertical Puts, Straddles, Strangles, Iron Condors, and a No-Trade action, covering the main dimensions of market uncertainty: directional exposure and volatility exposure. All positions are based on SPY options with a constant five-week (Friday-to-Friday) maturity, equivalent to approximately 35 calendar days to expiry and risk-defined at initiation, to ensure stable training and comparable reward scales, while single-leg options are excluded to avoid unbounded risk. These simplifications preserve payoff logic while keeping the action space compact. Each strategy family is represented by a single configuration: Verticals are debit spreads expressing bullish or bearish views, Iron Condors are credit spreads capturing short-volatility exposure, and Straddles and Strangles are debit trades seeking large moves. The underlying dataset consists of weekly SPY option chains between January 2020 and December 2024. Each record includes mid-quotes, bid–ask spreads, volume, open interest, and Black–Scholes Greeks (delta, gamma, vega, theta) computed using mid implied volatility. To incorporate broader context, macro and trend features are appended: the 20-day, 50-day, and 200-day moving averages of SPY, the VIX3M/VIX ratio (volatility-term slope), and the U.S. Treasury yield curve slope (10-year / 2-year). Liquidity filters retain only near-the-money, actively traded contracts with valid quotes, minimum open interest, and tight bid–ask spreads, ensuring realism and tradability. This limits the agent's exposure to mispriced opportunities that could be explored but

represents a conscious trade-off in favour of prioritizing stability over the exploitation of market inefficiencies. Each strategy candidate is matched with macro features that reflect regime context. Leg-level variables are aggregated into strategy-level feature vectors through simple averages, and the market features are supplemented with two derived measures: premium direction (credit versus debit) and distance to at-the-money (strike aggressiveness). The resulting tensors (i.e., multi-dimensional data arrays) encode normalized[1] features (X), tradability masks (M), and realized rewards (R) for every strategy candidate. This structure allows the PPO agent to operate on a stable and interpretable decision space.

Rewards correspond to the realized profit or loss that each strategy candidate would generate if held to expiry. Each position represents one SPY option spread (one lot, equivalent to 100 shares) and all rewards are expressed in USD per trade. Positions are not rebalanced or marked to market, and transaction costs and margin effects are excluded to maintain a uniform scale across actions. This design highlights structural profitability rather than short-term price dynamics. The advantage term (realized reward minus expected reward) drives policy updates, reinforcing decisions that outperform expectations. Additionally, the model follows a curriculum schedule that introduces strategy families sequentially: directional spreads first, then volatility-based structures, and finally condors, to allow for gradual learning adaptation or focus during training if needed. The dataset is organized chronologically, with the agent trained on the full 2020–2024 sample and evaluated on the most recent 20% days. This setup approximates forward-looking conditions and assesses the agent's ability to generalize within the historical distribution. The agent's performance was primarily assessed through the percentage of oracle performance, defined as the agent's average daily reward divided by a perfect-foresight benchmark. The oracle corresponds to the

[1] To avoid scale distortions, the features were subjected to z-score standardization, where each value was centered by subtracting the mean and scaled by dividing by the standard deviation.

best-performing tradable strategy within the same masked set on each day, providing a consistent upper bound for evaluating learning efficiency. This served as the main indicator of learning efficiency and guided model selection during training. Additional metrics such as average daily reward, Sharpe and Sortino ratios, hit rate, maximum drawdown, and the share of no-trade actions were then used to evaluate risk-adjusted returns, robustness, and trading discipline, capturing not only profitability but also the consistency and prudence of the learned trading behaviour.

To stabilise optimisation under noisy, fat-tailed payoffs, the process uses advantage normalisation, entropy regularisation, and gradient clipping. Entropy is gradually reduced to shift from exploration toward exploitation, guiding the agent toward consistent and economically coherent decision-making. Together, these design choices create a controlled but still realistic environment for the agent to learn stable, regime-specific trading logic.

6. Performance and Robustness

Framed by the architecture and evaluation framework described, the training process progressed through successive refinements of the PPO configuration[2]. Because RL depends on interaction and feedback, early runs displayed alternating phases of improvement and regression before stabilising around a consistent decision pattern. Key hyperparameters, including entropy decay and learning rate, were adjusted iteratively to improve convergence and prevent premature exploitation and training curriculum were selected in response to validation breakdown statistics of the strategies selected and performance. Multiple plateaus were encountered and overcome during this process. When performance, measured by the percentage of oracle, achieved more than 85% across repeated runs (possibly representing a plateau itself) the model was considered a valid checkpoint for detailed analysis. The

[2] Training was conducted through a custom interactive user interface that supported hyperparameter adjustment prior to each training iteration and monitored performance results after.

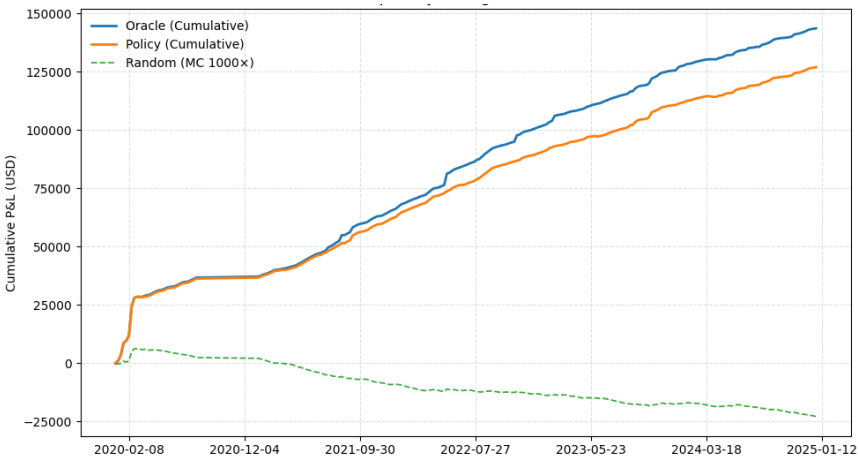
following table summarises the full set of metrics used to evaluate model performance at the selected checkpoint.

Table 1 – Policy performance summary at current checkpoint

Metric	Value
Mean daily reward	577 USD
% of oracle reward	88 %
Sharpe ratio	0.60
Sortino ratio[3]	2.76
Hit rate	95.9 %
Maximum drawdown	–378 USD
No-trade share	0.9 %

The validated policy demonstrates a disciplined and economically coherent trading style. It captures 88% of the oracle's attainable reward with moderate drawdowns and a hit rate near 96%, taking exposure only when market conditions appear favourable. This selective participation leads to smooth Sharpe and Sortino profiles and a small yet meaningful rate of no-trade actions. These confirm that the agent achieved coherent profitability while maintaining risk control and selective participation. It appears to favour modest but persistent gains rather than pursuing large, unstable payoffs.

Figure 1 – Cumulative rewards from Policy, Oracle, and Random Benchmark



[3] Since the environment focuses on absolute daily rewards rather than percentage returns, the Sortino ratio is computed as the average reward divided by the standard deviation of negative rewards.

Figure 1 shows that the policy stays between the oracle and random[4] benchmarks throughout the period. Its gains track the oracle closely early on, with a moderate gap emerging from late 2021, possibly reflecting a market shift to which the agent is less adapted. The overall tendency, steadiness, and clear separation from the random benchmark suggest that the policy reflects genuine learned pattern recognition.

Notwithstanding these results and being aware of the relatively small action space (in terms of the dataset's time span and volume of strategy candidates), the model's architecture is based on a flexible neural network updated directly from past rewards that can adapt too closely to historical data. It is therefore important to assess the possibility of overfitting. In this context, overfitting occurs when the agent learns to exploit specific historical patterns instead of developing a general rule that links market conditions to strategy outcomes. This would typically manifest as instability in performance over time, inconsistent actions under identical inputs, or profitability that fails to remain statistically significant once uncertainty and serial dependence are considered. Accordingly, to assess the checkpoint's temporal consistency and reliability, several robustness tests were performed, as described below. The corresponding methodologies and numerical results are presented in the Appendix.

I. A Cumulative Forward Window Test divided the dataset chronologically into cumulative windows covering the last 10%, 30%, and 50% of trading days. Performance remained stable across all windows at approximately 90% of the contemporaneous oracle, with improving Sharpe ratios, showing that the agent maintained temporal robustness and no evidence of overfitting to specific subperiods.

[4] The random benchmark was obtained through a Monte Carlo simulation, drawing one valid candidate uniformly per day from the same action set available to the policy (1,000 repetitions). The resulting daily rewards were averaged across simulations to form the expected random series. The simulation allowed the no-trade action, ensuring comparability with the policy's decision space.

II. A Rolling Window Test computed performance over overlapping 60-day windows advanced by 10 days to evaluate short-term temporal smoothness. All windows showed positive mean rewards and hit rates above 94%, confirming that profitability remained consistent through time and was not dependent on any particular subperiod.

III. A Sequential Window Test partitioned the full 220-day sample into five non overlapping forward windows of 44 days each to examine performance across distinct, consecutive market segments. All windows were profitable, with an average daily reward of approximately 470 USD and moderate dispersion, suggesting that variations in results reflected normal regime differences rather than model instability and supporting generalization capability.

IV. A Decision Stability Test re-queried the trained agent 100 times per day using identical market inputs to evaluate the determinism of policy behaviour. On average, the agent chose the same strategy 99% of the time, confirming the agent consistently selected the same strategy under identical conditions and converged toward a rule-like decision process.

V. A Moving Block Bootstrap Test applied a 15 day block bootstrap with 2000 resamples across 10%, 30%, and 50% tails of the reward series to assess statistical reliability while preserving serial correlation. The 95% confidence interval for the mean daily reward was approximately [410, 690], and the probability of a positive mean was 100%, confirming that profitability is statistically significant.

Overall, the robustness tests show that the PPO agent behaves as a stable, rule-based system with consistent and statistically significant profitability within the training sample, indicating that the model's logic stems from structured learning rather than random data fitting.

7. Regime Awareness

Having demonstrated the robustness of the model's performance, the analysis now turns to how the agent perceives market conditions and whether the patterns it recognises align with the regimes identified by humans. As discussed earlier, market regimes are recurring configurations of volatility, trend, and risk appetite that shape the performance of option strategies. While humans interpret these configurations through narrative and intuition, the PPO agent experiences them indirectly through the market features it observes and the realised payoffs that follow its actions. This raises the question of whether the agent organises its understanding of market conditions in a way that resembles human intuition, or whether it develops a distinct, data-driven perception of regimes.

Beyond its academic relevance, this analysis also serves a practical diagnostic purpose within the broader development of the RL framework. Understanding how the PPO agent perceives market structure provides a diagnostic view of its internal reasoning, revealing which features or relationships it has learned to prioritise. This insight can inform design refinements such as feature representation, reward design, and training curriculum.

In this context, the analysis below contrasts two independent segmentations of the same 2020–2024 SPY option environment: one representing a human benchmark derived from macro-volatility indicators, and another capturing the agent's representation inferred from its learned feature space. Both were identified using Principal Component Analysis (PCA) followed by K-means clustering[5]. Clustering is an unsupervised learning technique that groups observations with similar characteristics into distinct categories, allowing patterns of trend and volatility to emerge naturally and reveal how regimes are structured. Based on this approach, the 2020–2024 SPY market data clustered around six segmentations.

[5] Clustering was performed over a standardized feature space reduced through PCA, retaining 95% of total variance. The optimal segmentation was determined using the silhouette coefficient and the inertia-elbow criterion.

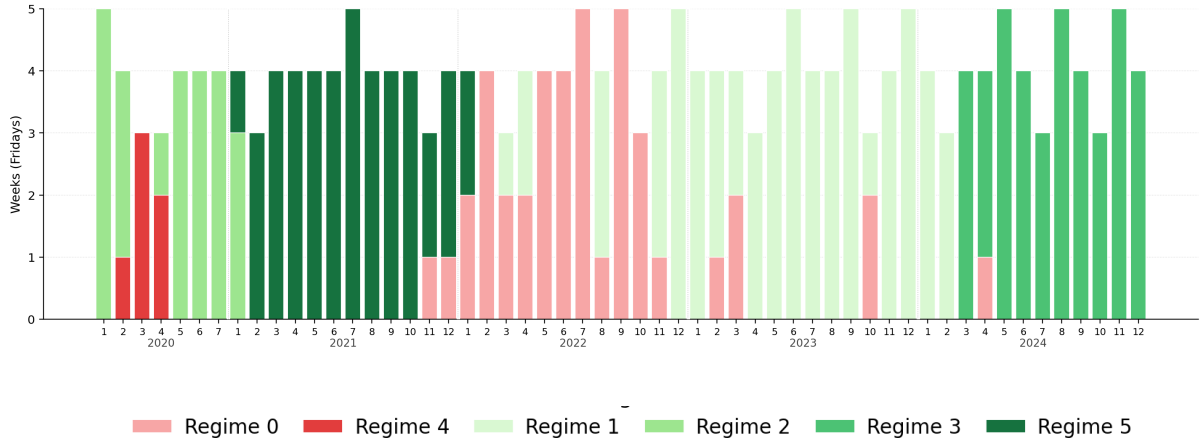
Table 2 – Results of the cluster analysis applied to market data (Cluster I)

Regime	Share of Days	SPX (20–200 MA)	VIX Level	VIX3M/VIX	10y–2y Slope	Mean daily reward		
						Policy	Oracle	% Oracle
0	18.6%	–3.8%	27.1	1.06	0.01	530	646	82.0%
1	29.6%	3.4%	18.9	1.16	–0.56	432	548	78.8%
2	10.9%	4.2%	26.0	1.10	0.46	1,465	1,466	99.9%
3	18.2%	9.3%	17.4	1.12	–0.13	338	362	93.4%
4	2.7%	–7.0%	52.0	0.89	0.43	386	453	85.2%
5	20.0%	10.1%	21.4	1.22	1.19	593	662	89.6%
Weighted Average						**577**	**653**	**88.4%**

These six regimes can be grouped into correction and expansion phases based on how volatility, yield-curve slope, and equity performance evolve. Regime 0 represents a moderate correction, where prices retreat modestly (SPX = –3.8%), volatility rises (VIX = 27%), and the yield curve flattens (0.01), signalling a short-term adjustment within an otherwise stable environment. Regime 4 reflects a stress marked with a sharp fall in prices (SPX = –7.0%), a surge in volatility (VIX = 52%), and a steepening curve (+0.43) as investors move toward safer short-term government bonds. As conditions stabilise, the market transitions into expansionary regimes. Regime 2 captures an early recovery, with prices rebounding (SPX = +4.2%), volatility easing (VIX = 26%), and a continued steep curve, reflecting expectations that growth will resume. Momentum strengthens in Regime 3, with solid gains (SPX = +9.3%) and low volatility (VIX = 17%), indicating a sustained expansion. This tendency peaks in Regime 5, which shows robust growth (SPX = +10.1%), low volatility (VIX = 21%), and a very steep curve (10y–2y = +1.19), consistent with abundant liquidity and market optimism. Finally, Regime 1 marks the late phase of the expansion, characterised by a smaller positive trend (SPX = +3.4%), low volatility (VIX = 19%), and an inverted yield curve (–0.56) that signals tightening financial conditions and slowing growth expectations.

The agent's efficiency relative to the oracle is highest in Regime 1, 2, 3, and 5 (around 90% to 100%), which together represent the predominantly positive, trend-following environments of the sample period. In contrast, efficiency stays at around 80% in Regimes 0 and 4, corresponding to corrections and stress episodes with higher volatility. This aligns with the calm and expansionary market context of 2020–2024, covering more than three quarters of trading days, with low realized volatility (17–26 %) and a persistently positive volatility-term slope (VIX3M/VIX > 1).

Figure 2 – Weekly regime characterization



The regime sequence mirrors the pandemic cycle: 2020 COVID-19 crash (Regime 4), shifting into early recovery (Regime 2) as stimulus took hold, then robust expansion through 2021's reopening boom (Regime 5). Tightening in 2022 brought correction (Regime 0) and then slight pickup in 2023 (Regime 1). Eventually markets settled into a solid expansion during 2024 (Regime 3). Consequently, any agent trained chronologically on this data will encounter far more short-volatility than long-volatility episodes. The agent's strategy mix per regime shows how this breakdown shaped its learned behaviour.

Table 3 – Agent's strategy mix by regime

Regime	Iron Condor	Vertical Call	Vertical Put	Straddle	Strangle	No-Trade
0	56%	15%	22%	0%	5%	2%
1	63%	17%	15%	0%	5%	0%
2	62%	4%	8%	17%	8%	0%
3	68%	22%	5%	0%	2%	2%
4	0%	50%	17%	33%	0%	0%
5	86%	0%	11%	2%	0%	0%

In the expansionary regimes (1, 2, 3, and 5), the agent favours Iron Condors capturing short-volatility exposure from their net credit at initiation. In the corrective regimes (0 and 4), the mix shifts toward Vertical Spreads and Straddles, reflecting a tactical adjustment towards volatility-buying strategies as the risk-reward of premium selling deteriorates and the exposure to market move compensate the upfront debit.

The dominance of positive regimes and the agent's higher efficiency in those markets show that it is regime-aware but less capable of exploiting volatility-buying opportunities. It learned to profit from stable, income-generating strategies but remains narrow and less responsive when volatility rises.

Next, to reveal how the agent organises its own perception of market states, we applied the same PCA + K-means framework to the feature tensors observed by the agent.

Table 4 – Results of the cluster analysis applied to agent's tensors (Cluster II)

Regime	Share of Days	Iron Condor	Vertical Call	Vertical Put	Straddle	Strangle	No-Trade
0	94.1%	70%	13%	12%	2%	3%	0%
1	5.9%	0%	23%	31%	15%	15%	15%

Rather than reproducing six regimes, the agent's feature space clusters into two distinct and stable behavioural modes, mirroring the distinction described above. Roughly 94% of trading days fall within a short-volatility income regime, dominated by Iron Condors. The remaining 6% correspond to a long-volatility or hedging regime, where the agent turns to Verticals,

Straddles, Strangles or No-Trade positions when volatility rises. This compression is not an oversimplification but an adaptation to market structure that reinforces the learning opportunities it was trained on.

Future design improvements should address this imbalance. Regime-balanced sampling could increase the weight of volatile or corrective periods, helping the agent learn from rare but impactful stress events. Adding temporal continuity rather than isolated episodes (i.e. Fridays only) would allow the agent to recognise transitions between regimes and anticipate volatility shifts instead of reacting to them. A regime-aware reward function could reduce incentives for short-volatility positions and/or reward effective long-volatility positions. Also, using a rolling retraining process that retains a curated "stress memory," could prevent the agent from forgetting defensive or long-volatility behaviours during extended calm phases. These refinements could move the framework to an even more adaptive trading system capable of maintaining profitability across evolving regimes.

8. Conclusion

This project explored RL as a means to *learn trading logic*, not to predict markets. By framing learning itself as a source of alpha, the study treated the PPO agent as a laboratory for decision-making and a way to observe how profitable behaviour can emerge from experience alone. Through repeated interaction with an options trading environment, the agent learned when to take risk, when to stay defensive, and when to do nothing. In doing so, it developed a behavioural pattern that mirrors how professional traders navigate market regimes: selling volatility when conditions are calm and shifting to long-volatility or neutral positions when uncertainty rises. The central achievement is not a model that "beats the market," but one that behaves coherently. The PPO agent has demonstrated that RL can internalize financially meaningful structure without external supervision. Its decisions are stable, interpretable, and economically consistent. That is significant from both a practical

and academic standpoint: it shows that a PPO framework can translate complex market feedback into actionable, rule-like behaviour, bridging the gap between quantitative automation and trader intuition. Yet this is only a checkpoint. The model remains deliberately simplified (single asset, fixed maturity, no execution costs) because the goal was understanding, not optimization. The next phase will focus on evolving this prototype into a live decision-support system capable of continuous learning. Each new episode of training adds to the agent's understanding of market structure, and each iteration offers us deeper insight into how trading rules adapt and fail. In this sense, the process of building and training the agent is itself a form of experiential learning. Future work will build on this foundation. With the guiding goal of achieving consistent and repeatable profitability, expanding into multiple assets is less valuable than refining the current SPY framework until it performs reliably under real trading conditions. The immediate priority is to bridge the gap between simulation and execution by introducing transaction costs, slippage and margin considerations. Additional training and evaluation cycles using fresh data and a larger dataset will be necessary to confirm stability across out of sample market regimes before committing real capital. Enhancing realism, and validating robustness through continued experimentation will mature the agent into a practical trading tool capable of sustaining performance without unnecessary complexity. Ultimately, this checkpoint proves that a disciplined, context-aware system capable of learning through experience is within reach.

9. References

Chujoy, Carlos, Joy Seth, and Satitpong Chantarajirawong. 2016. "*Use of Option Strategies to Improve Risk-Adjusted Returns on a 60/40 Investment Portfolio*." *CBOE Research Paper*. https://cdn.cboe.com/resources/education/research_publications/research-paper.pdf.

Liu, Lichan, Qing Li, and Siqi Fan. 2025. "The Impact of Volatility Regime Dynamics on Option Pricing." *North American Journal of Economics and Finance* 76: 102352. https://doi.org/10.1016/j.najef.2024.102352.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "*Proximal Policy Optimization Algorithms*." *arXiv preprint* arXiv:1707.06347. https://arxiv.org/abs/1707.06347.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction.* 2nd ed. Cambridge, MA: MIT Press. https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf.

Terven, Juan. 2025. "*Deep Reinforcement Learning: A Chronological Overview and Methods*." *AI* 6 (3): 46. https://doi.org/10.3390/ai6030046.

10. Glossary

Term	Definition / Formula
Actor–Critic	RL architecture with a policy network (actor) that selects actions and a value network (critic) that estimates expected returns to stabilise learning.
Advantage	Difference between realised and expected rewards used to update the policy: A(t) = R(t) − V(s(t))
Agent	Decision-making component that learns a policy mapping observed market features (state) to trading actions.
Bootstrap (Moving-Block)	Time-series resampling that draws consecutive blocks with replacement to preserve autocorrelation when estimating the distribution of a statistic (e.g., mean reward, Sharpe).
Contango / Backwardation	Slope of volatility term structure. Contango: longer-dated implied vols > shorter-dated (VIX3M/VIX > 1). Backwardation: the reverse.
Curriculum	Phased training schedule introducing strategy families sequentially.
Decision Stability Test	Re-query the trained agent multiple times with identical inputs.
Episode	One trading day in the environment. The agent selects a strategy at the start and receives a reward at expiry.
Greeks	Option sensitivities: Delta (price), Gamma (curvature), Theta (time decay), Vega (volatility).
Hit Rate	Share of profitable days: Hit Rate = profitable days / total days
Iron Condor / Straddle / Strangle / Vertical Call / Vertical Put	Standardised, risk-defined option structures combining legs to express short-/long-volatility and directional exposures.
Learning Rate	Step size controlling the magnitude of parameter updates during optimisation.
Market Regime	Recurring configuration of volatility, direction, and macro conditions influencing strategy performance.
Mask (Tradability Mask)	Binary filter restricting available actions to liquid, tradable strategy candidates.
Maximum Drawdown	Largest peak-to-trough loss in cumulative reward: = Maximum Drawdown = maximum(peak − lowest point that follows that peak)
Mean Daily Reward	Average realised profit or loss per trading day, in USD per one-lot position.
No-Trade Action	Explicit decision to take no position when expected reward is low or uncertainty high.
Oracle Benchmark	Perfect-hindsight chooser of the best tradable strategy each day from the same masked set. Represents the upper bound for learning efficiency.
% of Oracle	Efficiency Benchmark: % Oracle = 100 * mean(reward agent) / mean(reward oracle).
PCA (Principal Component Analysis)	Linear dimensionality reduction: transforms correlated variables into orthogonal components ordered by explained variance.
K-Means Clustering	Unsupervised partition into K groups by minimising within-cluster variance.
Policy	Probability distribution over actions learned by the agent and parameterised by the policy network.

Term	Definition / Formula
Proximal Policy Optimisation (PPO)	Policy-gradient RL with clipped update for stability.
Reward	Realised profit or loss (Rt) in USD per trade at expiry. It's the single learning signal for the agent.
Rolling Window Test	Performance evaluated on overlapping subperiods to assess temporal stability.
Forward Window Test	Chronological evaluation on cumulative forward segments to approximate out-of-sample behaviour.
Sharpe Ratio	Risk-adjusted return per unit of volatility: Sharpe = mean(R) / sd(R)
Sortino Ratio	Return per unit of downside volatility: Sortino = mean(R) / sd_down(R), where sd_down is the standard deviation of negative rewards only.
Strategy	Predefined option structure (verticals, straddles, strangles, condors, or no-trade) representing a discrete action.
VIX3M/VIX Ratio	Ratio of 3-month to 1-month implied volatility indices. Used as proxy for slope of the volatility term structure.
10y–2y Yield-Curve Slope	Ten-year minus two-year US Treasury yields. Used as indicator of macro expectations and risk appetite.