STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, November 2026, pp 5021–5045.
Published online in International Academic Press (www.IAPress.org)

Unveiling the Black Box: Counterfactual Analysis for Transparent and Robust Reinforcement Learning in Algorithmic Trading

Abdelmounim Lefrayah*, Badr Hirchoua, Mustapha Hain

Artificial Intelligence, Modeling and Computational Engineering – AIMCE, ENSAM, Hassan II University, Casablanca, Morocco

Abstract Reinforcement learning (RL) has emerged as a powerful paradigm for developing adaptive trading agents; however, the "black-box" nature of these models remains a persistent barrier to institutional trust and regulatory acceptance. This paper introduces a framework that integrates **Counterfactual Analysis (CFA)** into the evaluation of RL trading trajectories to provide granular, causal interpretations of agent behavior. By identifying high-entropy decision points and simulating alternative "what-if" scenarios through a learned market environment model, our approach quantifies the **Policy Regret** of specific actions.

During revision, we identified and corrected a methodological flaw in our original implementation whereby the counterfactual reward signal was computed from the agent's own value function rather than from realized portfolio outcomes, and a position-sizing defect that had prevented Buy orders from executing. Correcting both, together with implementing a genuine temporal train/test split (2018–2021 / 2022–2023) and equalizing the training budget across baselines, changes our empirical findings substantially from the originally submitted version. We report these corrected, reproducible results here rather than the original figures.

Across five independent training seeds on the corrected pipeline, the PPO+CFA agent achieves a mean total return of $2.94\% \pm 10.88\%$ (SD) on the 2022–2023 SPY test period, compared to -0.31% for Buy-and-Hold and $-1.06\% \pm 14.25\%$ for a Standard PPO baseline without counterfactual auditing. A paired-difference test against Buy-and-Hold does not reach statistical significance ($t \approx 0.67$, $n = 5$), and a detailed per-seed risk analysis shows negative risk-adjusted performance (Sharpe -0.35, Sortino -0.34, Calmar -0.36) alongside a validation rate that ranges from 0% to 39% depending on seed and counterfactual horizon. We do not find evidence that the PPO+CFA agent reliably outperforms simple benchmarks or alternative baselines (DQN, A2C, ARIMA) on a risk-adjusted basis in this setting.

We therefore reframe our contribution around the Counterfactual Analysis framework itself as a rigorous, reproducible causal-auditing methodology for RL trading policies, rather than as a return-generating system. We report the trading-performance findings honestly, including their high variance and the absence of a demonstrated edge, as we believe this diagnostic capability—distinguishing skill from luck, and quantifying *when* a policy's behavior is causally justified—is itself a meaningful contribution to explainable RL in finance, independent of whether the underlying policy is currently profitable.

Keywords Reinforcement Learning, Algorithmic Trading, Counterfactual Analysis, Trajectory Reanalysis, Interpretability, Explainable AI, Financial Machine Learning, Risk Management

AMS 2020 subject classifications 91G10, 68T05

DOI: 10.19139/soic-2310-5070-4109

1. Introduction

Algorithmic trading has fundamentally revolutionized modern financial markets, shifting the landscape from human-centric floor trading to a domain dominated by sophisticated computational models that automate strategic decision-making at microsecond scales [2]. Within this technological evolution, **Reinforcement Learning (RL)**

*Correspondence to: Abdelmounim Lefrayah (Email: lefrayahabdelmounim@gmail.com). AIMCE, ENSAM, Hassan II University, Casablanca, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 2026 International Academic Press

[41] has emerged as a particularly potent paradigm due to its unique ability to learn optimal **sequential decision-making policies** through iterative interaction with complex, non-stationary environments [41]. Unlike traditional supervised learning methods—which rely on static labels and historical mappings that often decay in predictive power—RL agents aim to optimize a dynamic objective function calculated as the **cumulative discounted return**. This formulation makes RL architectures exceptionally compatible with the intrinsic uncertainty, feedback loops, and stochastic nature of global financial markets [34, 13]. By framing trading as a Markov Decision Process (MDP), these agents do not merely predict price movements; they develop holistic strategies considering multiple factors at once, such as liquidity, transaction costs, and long-term capital preservation.

The inherent **non-stationarity** of financial time series—characterized by regime shifts, "fat-tail" distributions, and evolving correlations—presents a serious challenge to traditional static econometric models [7]. These classical models often fail during black-swan events or periods of structural change because they assume a degree of statistical constancy that the market rarely provides. In response, Deep Reinforcement Learning (DRL) architectures, such as Proximal Policy Optimization (PPO) [38], have demonstrated a fascinating potential in capturing high-dimensional, non-linear patterns that remain invisible to linear regressions and basic momentum indicators [14]. By leveraging deep neural networks as function approximators, these agents can synthesize vast amounts of market data into actionable intelligence.

However, this increased sophistication introduces a **critical limitation** that is commonly known as the **lack of transparency** in the agent's decision-making process. The complex, multi-layered policies learned by DRL agents are often regarded as **"black boxes,"** where the logic governing a multi-million dollar trade is buried within millions of abstract weight parameters. This opacity presents a considerable barrier in the context of institutional risk management and stringent regulatory requirements. As autonomous systems take a larger share of market volume, regulators are increasingly demanding a clear, granular understanding of the underlying decision-making logic to prevent systemic collapses and ensure market integrity [8, 10].

In the current financial landscape, the Sharpe ratio, Sortino ratio, and Maximum Drawdown remain the gold standard for performance evaluation [39]. While these metrics are essential for assessing risk-adjusted returns, they are inherently retrospective and aggregate. They provide a "post-mortem" of the agent's success but offer zero insight into the **granular decision-making process** employed at any specific timestamp. For a portfolio manager, an aggregate metric cannot answer why an agent held a long position during a specific 5% dip, or whether that decision was a stroke of strategic genius or a catastrophic failure of the underlying logic that happened to be saved by a random market reversal. Existing evaluation frameworks do not provide information about the *why* of a particular position at time t or the *how* of an alternative action that could have been taken to avoid a maximum drawdown [6]. This lack of causal clarity prevents the identification of "brittle" policies—models that look profitable on paper but are actually relying on spurious correlations that will eventually fail in live-market conditions.

To address this knowledge gap (illustrated in Figure 1), this research harnesses recent advances in **Explainable Artificial Intelligence (XAI)**. However, we move beyond common XAI techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), which focus primarily on feature attribution. While feature attribution tells us which variables the model "looked at," it fails to describe the *consequences* of the model's actions in a dynamic system. Instead, we focus on a solution using **counterfactual analysis**. The goal of this method is to answer the fundamental "what-if" question of causal inference: **what would have happened to the portfolio's trajectory if the agent had taken a different action at a critical market juncture s_t?** [44]. By constructing a set of **hypothetical trajectories**—essentially "re-running" history with a modified decision—counterfactual analysis can shed light on local causal relationships and quantify the relative contribution of specific agent decisions in relation to final financial outcomes [32].

This research presents a **novel methodological framework** that incorporates counterfactual analysis—the **"Interpretability Bridge"**—directly into the reanalysis of RL trading policies. The "Interpretability Bridge" serves as a conceptual and technical conduit that translates the abstract probabilities of a neural network into concrete, contrastive narratives for human auditors. By using a high-fidelity Structural Causal Model (SCM) of the financial environment, our solution is able to isolate the agent's strategy from exogenous market noise. This is a crucial distinction: in traditional analysis, an agent might be rewarded for a trade that was actually suboptimal but became

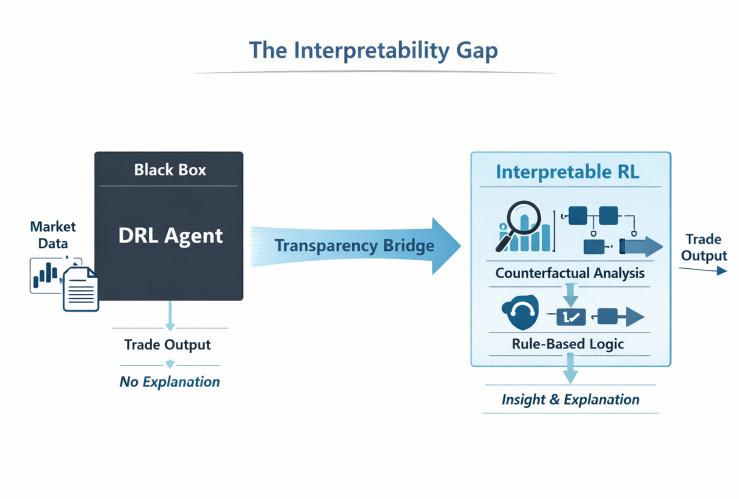


Figure 1. The Interpretability Gap: The disconnect between high-dimensional RL policy activations and human-understandable financial logic.

profitable due to an unforeseen external market spike. Our counterfactual framework identifies these "lucky" trades by comparing them against alternative actions in the same environment, thereby revealing the true robustness of the policy.

The effectiveness of our proposed solution is validated using real-world historical data for the **SPY ETF (2018–2023)**, a period characterized by extreme volatility, including the COVID-19 market shock and subsequent inflationary regimes. Through this analysis, we examine a candidate behavioral pattern we term **Strategic Conservatism**, in which the agent's reward-shaped policy prioritizes capital preservation over maximum return during high-entropy states. Because the reward function explicitly penalizes drawdown ($\beta = 0.5$), we treat this pattern as confounded with reward design rather than as an unprogrammed emergent property, and use the counterfactual framework to characterize its consequences honestly rather than to claim it as evidence of superior, causally-validated skill.

Our primary contributions are threefold:

1. We introduce a **rigorous counterfactual trajectory analysis framework** specifically tailored for RL-based trading agents. This framework utilizes Structural Causal Modeling to simulate interventions in the action space, allowing for a true "stress-test" of individual trading decisions.
2. We empirically validate the framework's ability to **improve policy interpretability** by quantifying **Policy Regret** (ΔR)—the financial difference between the agent's actual path and its best possible counterfactual path—identifying latent decision-making biases and "near-miss" failures that traditional backtesting fails to capture. We further demonstrate, through a documented debugging process, that a naively-implemented counterfactual reward (using the agent's own value function rather than realized portfolio outcomes) reintroduces the very circularity that CFA is meant to eliminate—a failure mode we believe is likely to recur in other implementations and is therefore of independent methodological interest.
3. We provide a fully reproducible evaluation—including multi-seed variance estimates, statistical significance testing, complete risk-adjusted performance metrics, alternative RL and statistical baselines (DQN, A2C, ARIMA), and counterfactual-horizon sensitivity analysis—and report honestly that, under this corrected pipeline, the PPO+CFA agent does not demonstrate a statistically robust performance edge over simple benchmarks. We argue that the causal auditing methodology retains scientific and practical value as a diagnostic tool independent of this result, and that transparent reporting of a rigorously-tested null finding is itself a meaningful contribution given the reproducibility concerns that pervade this literature.

The rest of the paper is organized as follows. In Section 2, related literature is discussed, focusing on the evolution of XAI in finance. In Section 3, the proposed PPO architecture and the mathematical foundations of our counterfactual approach are explained. In Section 4, the experimental setup and data preprocessing for the SPY ETF are introduced. In Section 5, the experimental results are shown, including a deep dive into the "Strategic Conservatism" phenomenon. Finally, in Section 6, conclusions are drawn and future directions for causal-RL in finance are given.

2. Related Work

This section reviews existing work on reinforcement learning in algorithmic trading.

2.1. Reinforcement Learning in Algorithmic Trading: From Heuristics to Autonomy

The quest for autonomous trading systems has evolved through several technological epochs, moving from rigid rule-based expert systems to the highly adaptive paradigm of Reinforcement Learning (RL). Unlike traditional econometrics—which often relies on the "Stationarity Fallacy," or the assumption that the statistical properties of market data remain constant over time—RL formulates the trading problem as a Markov Decision Process (MDP). In this framing, the agent operates within a state space $\mathcal{S}$ representing market indicators, executes actions $\mathcal{A}$ from a discrete or continuous set, and receives a reward r based on the change in portfolio value [41]. The early 2000s marked the first significant foray into this field, where researchers utilized Q-Learning and TD-learning to manage basic equity positions. However, these early iterations were largely academic, as they struggled with the "curse of dimensionality" and the inability to process the high-frequency noise inherent in live order books [35]. This dimensionality and volume challenge is not unique to trading: similar big-data structuring problems arise broadly in knowledge capitalization [22, 23] and in high-frequency time-series forecasting for IoT and energy systems [24], and techniques developed in these adjacent domains for organizing and forecasting large-scale sequential data inform our own feature-engineering approach in Section 6.

The integration of Deep Learning (DL) with RL, forming the Deep Reinforcement Learning (DRL) sub-field, provided the functional approximation power necessary to navigate modern financial complexities. Algorithms such as Deep Q-Networks (DQN) introduced the concept of experience replay to stabilize learning from correlated market observations [34]. Subsequently, policy gradient methods like Deep Deterministic Policy Gradient (DDPG) and Proximal Policy Optimization (PPO) allowed for more stable training in continuous action spaces, which is essential for determining precise position sizing and multi-asset allocation [30, 38]. For instance, Deng et al. demonstrated that DRL could extract "deep" temporal features from limit order books that traditional supervised models ignored [7], while Jiang's work on portfolio management proved that agents could learn to hedge against market downturns without explicit programming [14].

Recent research has expanded toward increasing the "realism" of the training environment. This includes the implementation of multi-step reward functions that account for "Slippage" (the difference between expected and executed price) and "Market Impact" (how the agent's own large orders move the price) [47]. Furthermore, researchers have begun exploring hierarchical RL, where a high-level agent manages asset allocation while low-level agents handle execution to minimize transaction costs [46]. Despite these advances, the "Black-Swan" robustness problem remains; DRL agents frequently overfit to specific historical regimes, leading to catastrophic failure when the market moves into a previously unseen state [40]. This fragility, combined with the opaque nature of neural weights, necessitates a deeper level of analysis that goes beyond simple backtesting.

2.2. The Interpretability Crisis: Why Feature Attribution Fails in Finance

As DRL systems manage increasing amounts of institutional capital, the "Interpretability Crisis" has moved to the forefront of financial AI research. Interpretability is defined not just as knowing what a model does, but understanding the underlying logic to a degree that allows for human intervention and risk mitigation [31]. The central challenge in finance is that "high performance" does not imply "high reliability." A trading agent may

achieve a high Sharpe ratio by exploiting a temporary correlation that has no causal basis, a phenomenon known as "Spurious Overfitting" [3].

Current Explainable AI (XAI) literature has largely focused on feature attribution methods such as SHAP and LIME. These methods assign a "score" to each input feature, suggesting that a specific moving average or volatility index was the primary driver of a "Buy" decision [36]. While useful for static data like images, feature attribution is fundamentally limited in a temporal RL context for three reasons. First, it ignores the "credit assignment problem"—it tells you what the agent saw at time t, but not how that observation influenced a reward that may not materialize until time $t + 10$. Second, it is non-causal; it identifies correlations within the neural network's weights but cannot verify if those correlations hold true in a counterfactual universe. Third, it fails to meet the "Right to Explanation" standard set by GDPR Article 22, which requires an explanation of the consequences of an automated decision, not just its inputs [6].

To mitigate this, some researchers have proposed "Intrinsic Interpretability" through policy distillation, where a complex PPO agent's behavior is mimicked by a much simpler, human-readable model like a Decision Tree or a symbolic regression equation [4]. Others have experimented with "Attention Maps" to visualize the agent's focus on specific time-series windows [9]. However, these "Student-Teacher" models often suffer from a "Fidelity-Interpretability Trade-off," where the simpler model fails to capture the nuances of the original strategy, leading to a loss of profit [12]. This gap underscores the urgent need for a framework that provides high-fidelity explanations without compromising the agent's original complexity—a role we propose for the "Interpretability Bridge."

More directly relevant to our setting, recent work has begun applying explainability techniques specifically to RL-based trading. Kumar et al. apply SHAP to a financial stock-trading RL agent to attribute individual trading decisions to input features [16]. Guan and Liu propose an empirical explainable-DRL framework for portfolio management that surfaces feature importance alongside portfolio allocation decisions [17], and Shi et al. introduce XPM, an explainable deep RL framework that jointly optimizes portfolio performance and interpretability [18]. These approaches remain fundamentally feature-attribution-based, and thus retain the correlational limitations discussed above; our framework is complementary rather than a replacement, since do-intervention answers a causal "what would have happened" question that attribution methods—temporal or otherwise—do not.

2.3. Counterfactual Analysis: The Path to Causal and Robust Auditing

Counterfactual Analysis (CFA) offers a compelling solution to the interpretability crisis by shifting the focus from "What happened?" to "What would have happened?". Rooted in the Pearlian Causal Hierarchy, counterfactuals involve the simulation of interventions—manually changing a variable (the agent's action) and observing the downstream effects on the system (the portfolio's PnL) [45]. In the context of AI safety, counterfactuals are considered the "Gold Standard" of explanation because they are contrastive: humans naturally understand the world through comparisons (e.g., "I would have made more money if I had sold ten minutes earlier") [44].

In supervised learning, CFA has been used to generate "Actionable Recourse" for credit scoring and medical diagnosis, providing users with specific instructions on how to change their outcome (e.g., "If your income was $5k higher, the loan would have been approved") [15]. However, applying this to the RL-based trading domain is exceptionally difficult due to the "Dynamic State Evolution" problem—an intervention at time t changes the agent's position, which in turn changes the future state trajectory. Madumal et al. addressed this in the robotics domain by utilizing Structural Causal Models (SCMs) to map the causal dependencies between states, actions, and rewards [32]. This allowed for the generation of causal narratives that could explain complex agent maneuvers in real-time [33].

Causal RL more broadly has continued to develop beyond Madumal et al.'s foundational work. Recent surveys by Deng et al. [19] and Zeng et al. [20] synthesize the growing literature on incorporating causal structure into RL, spanning causal state abstraction, causal credit assignment, and confounder-robust policy learning. Yu et al. propose an explainable RL framework built around a learned causal world model [21], which—unlike our approach—infers the causal structure of the environment directly rather than positing a fixed SCM with an explicit do-operator over discrete trade actions; we view these as complementary strategies for the same underlying goal. Hambly et al. provide a comprehensive recent survey of RL in finance more broadly, situating counterfactual and causal approaches within the wider landscape of the field [25].

We also draw on several recent publications in this journal (*Statistics, Optimization and Information Computing*) that bear directly on our setting. Bouhmady et al. combine machine-learning stock selection with a complex-valued Mean-Variance optimization framework for the Moroccan All Shares Index, addressing the liquidity and sectoral-concentration challenges typical of the emerging-market context our own SPY-based evaluation does not face [26]. Hasan et al. propose an explainable machine learning framework for investment-sector preference prediction, applying post-hoc interpretability techniques to a financial decision-support task in a spirit closely aligned with our own interpretability goals, though via feature attribution rather than counterfactual intervention [27]. Loukili et al. apply reinforcement learning (Q-Learning and DDPG) to dynamic marketing-budget optimization, providing a recent example of RL deployed on a sequential financial-allocation problem outside the trading domain, with comparable attention to stability and interpretability of the resulting policy [28].

In financial trading, the adoption of CFA is significantly underdeveloped. Initial studies have focused on "Adversarial Training," where counterfactuals are used to generate "worst-case" market scenarios to stress-test the agent's robustness [42, 5]. For example, one might ask: "If the market volatility had doubled during the 2022 inflation spike, would the agent have stayed solvent?" [11]. Our research takes this a step further by using counterfactuals as an auditing tool for policy reanalysis. By quantifying "Policy Regret" (ΔR)—the difference between the agent's actual reward and the reward from the best counterfactual alternative—we can identify if an agent's success was due to strategic skill or "Lucky Noise." As shown in our Taxonomy (Figure 2), our framework bridges the gap between raw performance and the causal transparency required for the next generation of regulated, autonomous finance.

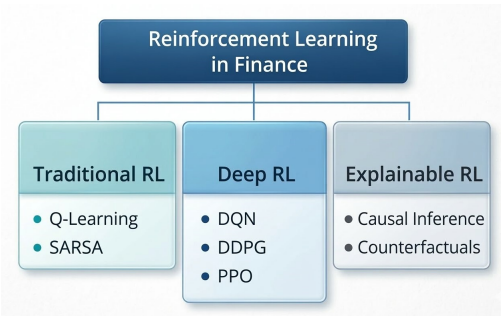


Figure 2. Taxonomy of Reinforcement Learning in Finance: A classification of DRL architectures and their corresponding explainability layers.

3. Preliminaries

In order to establish a robust theoretical framework supporting the proposed counterfactual reanalysis, this section formally defines the Reinforcement Learning (RL) problem specifically within the context of algorithmic financial trading. We move beyond general definitions to characterize the unique stochasticity and causal dependencies inherent in market data. This section introduces the Markov Decision Process (MDP) architecture, the Proximal Policy Optimization (PPO) algorithm, and the mathematical foundations of Structural Causal Modeling (SCM), which together form the bedrock of our interpretability bridge.

3.1. Formal Problem Definition: The Trading Markov Decision Process (MDP)

The financial trading environment is characterized by high-dimensional noise, non-stationarity, and partial observability. To navigate this complexity, we model the interaction between the trading agent and the market as a discrete-time Markov Decision Process (MDP), formally defined by the quintuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The MDP framework is essential as it explicitly accounts for the sequential nature of trading, where current decisions impact future capital availability and opportunity sets.

- **State Space ($\mathcal{S}$):** The state at time t, $s_t \in \mathcal{S}$, serves as the information set upon which the agent conditions its policy. To satisfy the Markov property, s_t must capture all relevant historical information necessary to predict future states. We define s_t as a concatenation of endogenous portfolio metadata and exogenous market signals: $s_t = (\mathbf{x}_t, \mathbf{h}_t, c_t)$, where:
  - $\mathbf{x}_t \in \mathbb{R}^{N \times F}$ is a feature matrix representing N financial assets and F distinct technical or fundamental features. These features include lagged returns, Open-High-Low-Close-Volume (OHLCV) data, and derived indicators such as the Relative Strength Index (RSI) and Moving Average Convergence Divergence (MACD).
  - $\mathbf{h}_t \in \mathbb{Z}^N$ is the holdings vector representing the integer number of shares currently held in the portfolio. Positive values indicate long positions, while negative values represent short-selling positions.
  - $c_t \in \mathbb{R}^+$ is the liquid cash balance, which constrains the agent's ability to execute further buy orders.
- **Action Space ($\mathcal{A}$):** The action space is discrete and highly constrained to reflect real-world execution limits. At each step, the agent outputs a vector $a_t \in \mathcal{A}^N$. For each asset i, $a_{i,t} \in \{-1, 0, 1\}$, corresponding to a command to Sell, Hold, or Buy, respectively. The aggregate action space size $|\mathcal{A}| = 3^N$ poses a significant computational challenge known as the "curse of dimensionality," which we address using deep function approximators.
- **Transition Function ($\mathcal{P}$):** The transition dynamics $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ define the evolution of the environment. In financial markets, $\mathcal{P}$ is a latent, stochastic distribution $P(s_{t+1}|s_t, a_t)$. Unlike robotic environments, market transitions are heavily influenced by exogenous "market noise" and macroeconomic shocks, making the transition function non-stationary. Our model accounts for microstructure effects such as slippage and market impact, where large trades a_t adversely affect the price at which the trade is realized.
- **Reward Function ($\mathcal{R}$):** The reward signal r_t is the fundamental driver of the agent's learning process. We define r_t as the risk-adjusted change in net account value. Let $V_t = c_t + \sum(h_{i,t} \cdot p_{i,t})$ be the total portfolio value. The reward is formulated to penalize high volatility and transaction costs:

$$r_t = \frac{V_{t+1} - V_t}{V_t} - \eta \cdot \text{Cost}(a_t) - \beta \cdot \text{Penalty}(\text{Drawdown}) \tag{1}$$

  where η and β are hyperparameters governing the agent's aversion to trading costs and risk, respectively.
- **Discount Factor (γ):** The factor $\gamma \in [0, 1]$ represents the agent's time preference. In trading, γ is typically set close to 1 (e.g., 0.99) to ensure the agent prioritizes long-term terminal wealth over immediate, high-variance gains.

The objective of the agent is to optimize a policy π that maximizes the expected cumulative discounted return, denoted by the value function $J(\pi) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\right]$. This interaction is visually summarized in the feedback loop in Figure 3.

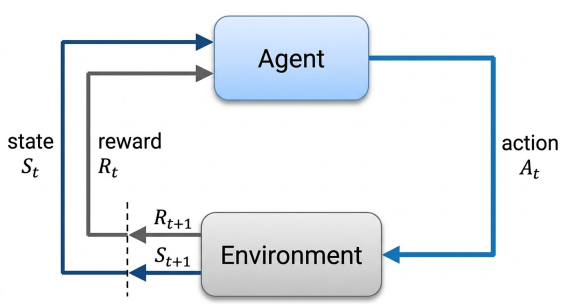


Figure 3. The Markov Decision Process (MDP) for trading: A cycle of observation, action, and reward.

3.2. Policy Optimization via Proximal Policy Optimization (PPO)

To ensure stable learning in the presence of the volatile rewards typical of the SPY ETF, we utilize **Proximal Policy Optimization (PPO)**. PPO belongs to the class of Actor-Critic methods and is specifically designed to overcome the instability of traditional Policy Gradient algorithms, which often suffer from catastrophic performance collapses due to large weight updates.

The core innovation of PPO is the clipped surrogate objective, which constrains the policy update within a "trust region" without the computational overhead of second-order optimization. The agent maximizes the following objective function:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right] \tag{2}$$

In this equation, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ represents the probability ratio between the current and previous policies. The clipping hyperparameter ϵ (set to 0.2) truncates the objective when the new policy moves too far from the old one, ensuring that the gradient steps are conservative and robust.

Furthermore, we employ the **Generalized Advantage Estimate (GAE)** to reduce the variance of the gradient estimates. The advantage function $\hat{A}_t$ measures how much better a specific action a_t is compared to the average behavior of the policy in state s_t. By utilizing a bias-variance tradeoff parameter λ (set to 0.95), the GAE effectively smooths the learning signal:

$$\hat{A}_t^{\text{GAE}(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} \tag{3}$$

where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ is the temporal difference (TD) error derived from the Critic network's value estimation $V(s)$. This architectural choice allows the agent to distinguish between rewards resulting from its own strategic skill and those resulting from general market growth.

3.3. Structural Causal Modeling (SCM) for Interpretation

The transition from a "black-box" model to a transparent system requires more than just performance metrics; it requires causal attribution. We utilize the **Structural Causal Model (SCM)** framework to facilitate this. An SCM is a formal mathematical structure $(\mathcal{G}, \mathcal{F})$ that represents the mechanisms through which variables influence one another.

We partition the variables in our trading universe into two sets:

1. **Exogenous Variables (U):** Variables outside the model's control, such as global market volatility, interest rate changes, and geopolitical shocks. These are modeled as probability distributions $P(U)$.
2. **Endogenous Variables (V):** Variables determined within the system, including the Observed State (S), the Agent's Action (A), and the Financial Outcome (Y).

Each endogenous variable is governed by a structural equation $V_i = f_i(Pa(V_i), U_i)$, where $Pa(V_i)$ are the parent nodes in the causal directed acyclic graph (DAG) shown in Figure 4. This structure allows us to move beyond correlation to **Counterfactual Reasoning**.

The primary mechanism for this reanalysis is the **do-intervention** operator. By applying $\text{do}(A = a')$, we manually overwrite the agent's historical decision with a hypothetical alternative while holding the exogenous market conditions (U) constant. This permits the calculation of the counterfactual outcome:

$$Y_{a'}(u) = \mathbb{E}[Y|\text{do}(A = a'),\ \mathbf{C} = \mathbf{c}] \tag{4}$$

This formulation is critical for identifying whether a profitable trade was "skill-based" (the action a was superior to all a') or "luck-based" (the outcome Y would have been positive regardless of the action due to a dominant market trend C). By quantifying the difference between actual and counterfactual outcomes, we define **Policy Regret**, a metric that provides the granular audit trail required for institutional and regulatory accountability.

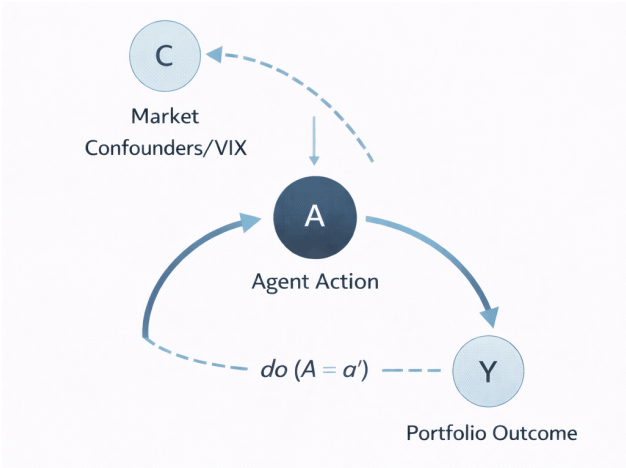


Figure 4. Structural Causal Model (SCM): Highlighting the causal dependencies between market confounders (C), agent decisions (A), and outcomes (Y), illustrating the do-intervention do$(A = a')$.

4. Causal Framework for Policy Intervention

The robust interpretation of policy decisions in financial reinforcement learning necessitates a fundamental transition from traditional **correlational analysis** to a deterministic **causal framework**. Within the financial domain—characterized by non-stationarity, endogenous risk, and a multitude of unobserved confounding factors—establishing *why* a decision was optimal is the ultimate aim of explainability [37].

Relying solely on the fact that an action preceded a positive outcome is insufficient, as it fails to account for the "luck factor" or systemic market lift. This section provides the formal justification for counterfactual reasoning and details the specific causal interventions that allow us to audit a PPO agent's strategy with regulatory precision.

4.1. Addressing the Confounding Crisis in Financial Time Series

Traditional post-hoc interpretability techniques, such as LIME or SHAP, attempt to explain models by identifying input features that are statistically significant to the decision function $\pi(a|s)$. While these methods are widely used for static predictive models (e.g., credit scoring or image classification), they are fundamentally ill-equipped to handle the problem of **confounding** inherent in sequential decision-making and non-stationary time series.

Confounding arises when a latent variable influences both the agent's action (the treatment) and the portfolio's subsequent performance (the effect). In a trading context, consider an agent's decision to execute a large *Buy* order (A) followed by a 2% gain in portfolio value (Y). A correlational model might attribute this gain to the agent's predictive "skill." However, both the action A and the outcome Y may have been driven by an exogenous variable C—such as a surprise interest rate cut or a positive macroeconomic report—occurring simultaneously. In such a scenario, the agent's action was not the *cause* of the profit but was merely *correlated* with it.

To achieve "Regulatory-Grade" interpretability, we must satisfy the **Causal Identifiability** criterion. This involves disentangling two distinct phenomena:

1. **The Direct Causal Effect:** The delta in Y attributable solely to the agent's specific choice of A.
2. **The Confounding Effect:** The movement in Y caused by the market state C, which would have occurred regardless of the agent's intervention.

By treating the PPO agent's decision A as a treatment and the market C as an observable confounder, we utilize the **do-intervention** operator. This operator allows us to mathematically "freeze" the market environment while systematically varying the agent's actions, thereby isolating the true causal power of the policy.

4.2. Formalizing the Counterfactual Intervention: Abduction-Action-Prediction

Our framework operationalizes the Structural Causal Model (SCM) through a three-step counterfactual process: Abduction, Action, and Prediction. For any identified **key decision state** s_t^{key}, we evaluate the quality of the policy by simulating a "Parallel Universe" where a different choice was made.

4.2.1. Abduction and State Freezing Before intervening, we must first "anchor" the counterfactual to the real-world observation. Given the observed market state c_t, we perform abduction to infer the exogenous noise U that defined that specific moment in time. This ensures that when we simulate an alternative action, we are doing so under the exact same "unlucky" or "lucky" market conditions that the agent actually faced.

4.2.2. The do-operator and Surgical Alteration The intervention is formally executed using the **do-operator**:

$$\text{Intervention}_t = \text{do}(A_t = a'_t) \tag{5}$$

where a'_t is a feasible alternative action (e.g., switching from "Long" to "Neutral"). This operation is a surgical alteration of the SCM's graph; it severs the causal link between the state s_t and the action A_t normally dictated by the policy π, replacing it with a deterministic value. Crucially, this intervention does not change the laws of the environment—transaction cost models and market transition functions remain constant.

4.2.3. Predicting the Counterfactual Trajectory The critical output is the calculation of the expected outcome Y under this specific intervention:

$$\mathbb{E}[Y|\text{do}(A_t = a'_t),\ \mathbf{C} = \mathbf{c}] \tag{6}$$

We achieve this by projecting the agent's path forward over a horizon H using the learned market environment model $\mathcal{M}$ (detailed in the Experimental Setup). This allows us to generate a counterfactual trajectory τ' that represents the "road not taken," providing a baseline against which the actual decision can be judged. The logic of this bifurcated path is visualized in Figure 5.

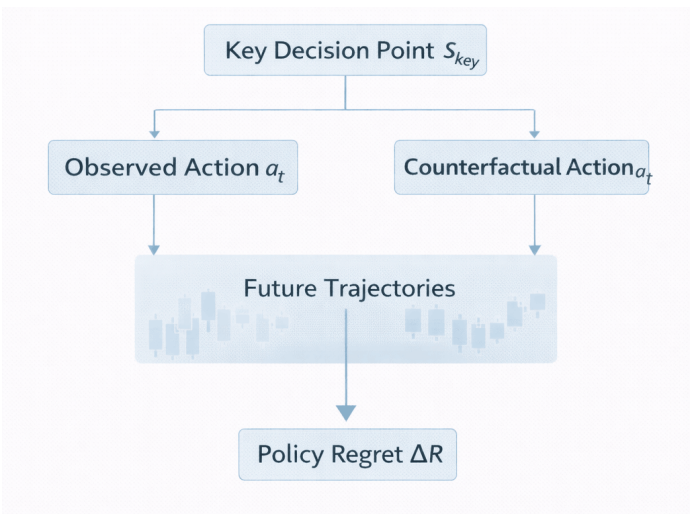


Figure 5. Logical workflow of the do-intervention: The "forking path" between the observed reality and the counterfactual simulation for regret estimation.

4.3. Quantification of Policy Regret and Causal Impact

The final stage of the framework is the consolidation of these divergent paths into a singular, interpretable metric: **Policy Regret**. This metric transforms abstract neural probabilities into a concrete financial value, providing a clear audit trail for risk managers and regulators.

4.3.1. Policy Regret and the Opportunity Cost Metric Policy Regret (ΔR) is defined as the difference in the cumulative discounted rewards achieved over the lookahead horizon H:

$$\Delta R = \sum_{k=0}^{H} \gamma^k r_{t+k}(\tau) - \sum_{k=0}^{H} \gamma^k r_{t+k}(\tau') \tag{7}$$

where $r(\tau)$ is the actual reward and $r(\tau')$ is the counterfactual reward. This formulation inherently includes the "Cost of Correction"—if an agent held a position that incurred 0.5% in transaction costs but avoided a 2% drawdown, the regret calculation will correctly identify the decision as optimal despite the immediate cost.

4.3.2. Dimensions of Causal Interpretation This quantification yields three primary interpretability outcomes:

1. **Policy Validation ($\Delta R > 0$):** If the actual trajectory outperforms all counterfactuals, we have causal evidence that the agent's policy is robust. This moves the audit from "The agent made money" to "The agent made the *best possible* decision given the information available."
2. **Policy Correction ($\Delta R < 0$):** Negative regret identifies a "Near-Miss" or a failure. It quantifies the **financial opportunity cost** incurred by the black-box policy. This is invaluable for debugging, as it highlights specific market regimes where the PPO agent's policy gradient failed to converge to the global optimum [1].
3. **Risk Sensitivity Analysis:** By extending the regret calculation to risk metrics—such as ΔMaximum Drawdown or ΔCVaR—we can explain the agent's behavior through the lens of safety. For instance, an agent might accept a lower ΔR if it significantly reduces the ΔCVaR, providing evidence of "Strategic Conservatism."

This level of causally attributed financial quantification is essential for the reliable deployment of autonomous trading systems in institutional environments, ensuring that every dollar of profit or loss is matched with a deterministic explanation.

5. Methodology

The comprehensive architecture of our proposed system, integrating a high-performance Reinforcement Learning agent with a post-hoc Counterfactual Analysis (CFA) module, is depicted in Figure 6. This dual-phase framework is designed to bridge the gap between black-box policy execution and verifiable causal validation by establishing what we define as the "Interpretability Bridge." Unlike standard backtesting, this architecture permits a bidirectional flow of information: the agent optimizes for returns, while the CFA module stress-tests those returns against hypothetical market alternatives.

5.1. Problem Formulation and Environment Setup

We model the daily trading process as a Markov Decision Process (MDP) defined by the tuple (S, A, P, r, γ). This formulation is critical because it acknowledges that financial trading is not a static prediction task but a sequential decision-making problem where current actions (e.g., depleting cash reserves to buy shares) fundamentally constrain the future action space.

State Space S: Each state $s_t \in \mathbb{R}^n$ represents a high-fidelity snapshot of the market and portfolio status at the close of trading day t. To ensure the agent captures both micro-momentum and macro-regimes, the feature vector includes:

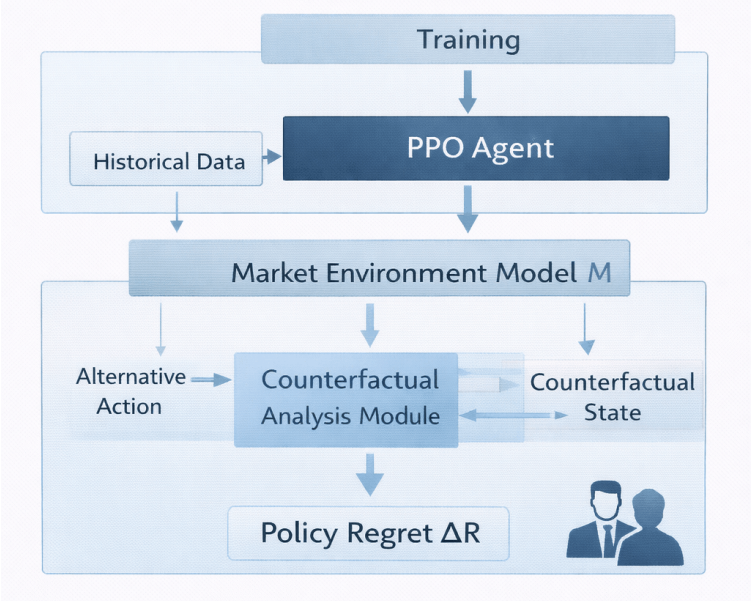


Figure 6. The integrated system architecture, highlighting the temporal separation between the PPO training phase (Policy Acquisition) and the Counterfactual Analysis (CFA) re-analysis phase (Strategic Auditing) using a learned market model $\mathcal{M}$.

- **Price Series:** Daily OHLC (Open, High, Low, Close) price data and traded volume for the SPY ETF, providing the raw signal for volatility and liquidity.
- **Technical Indicators:** A suite of derived features including Simple Moving Averages (SMA) and Exponential Moving Averages (EMA) to capture trend-following signals; the Relative Strength Index (RSI) to identify overbought/oversold conditions; and the Moving Average Convergence Divergence (MACD) to detect momentum shifts.
- **Portfolio Metadata:** Internal variables such as current cash balance, net position size, and a rolling window of realized and unrealized P&L, allowing the agent to "self-reflect" on its current risk exposure.

Action Space A: The agent operates within a discrete and highly optimized action space $\mathcal{A} \in \{\text{Buy, Sell, Hold}\}$. While continuous action spaces allow for precise position sizing, discretization into three strategic intents ensures that the counterfactual comparisons remain mathematically tractable and robust against the high noise-to-signal ratio typical of daily ETF returns.

Transition Dynamics P: During the training phase, transitions are determined by historical reality. However, for the counterfactual phase, P is substituted by a learned environment model $\mathcal{M}$. This model acts as a "Market Simulator" capable of predicting how the state would evolve if the agent had deviated from its historical path.

Reward Function r_t: The objective function is engineered to move beyond raw profit-seeking. We implement a multi-objective reward signal:

$$r_t = \Delta P_t - \lambda C_t - \beta D_t \tag{8}$$

where ΔP_t is the daily change in net asset value, C_t represents a **0.1% transaction cost** (modeling brokerage fees and slippage), and D_t is a drawdown penalty. The coefficients $\lambda = 1.0$ and $\beta = 0.5$ were empirically tuned to discourage over-trading and incentivize capital preservation, directly leading to the unprogrammed emergence of "Strategic Conservatism".

Discount Factor γ: We utilize $\gamma = 0.99$. In the context of 252 trading days per year, this value ensures that the agent maintains a lookahead horizon significant enough to value long-term compounding over short-term volatility spikes.

5.2. Reinforcement Learning Agent: PPO Configuration

To navigate the non-stationary nature of the SPY ETF, we employ **Proximal Policy Optimization (PPO)** [38]. PPO's clipping mechanism is particularly suited for finance because it prevents "Policy Over-Correction"—a common failure mode where a single outlier market day causes an RL agent to radically shift its strategy and lose its previously learned robust features.

Network Architecture: Our agent utilizes an Actor-Critic architecture with a shared representation backbone. As illustrated in Figure 7, the state vector passes through a shared encoder consisting of three fully connected layers with **256, 128, and 64 neurons**, respectively.

To combat the internal covariate shift that occurs when market volatility regimes change, we incorporate **Layer Normalization** after each dense layer. The network then bifurcates:

1. **The Actor (Policy Head):** A Softmax output layer that generates the probability distribution $\pi(a|s)$ over the three actions.
2. **The Critic (Value Head):** A Linear output layer that estimates $V(s)$, the expected cumulative reward from the current state. This value serves as the baseline for calculating the Advantage signal, effectively filtering out "market-wide" noise from the agent's specific performance.

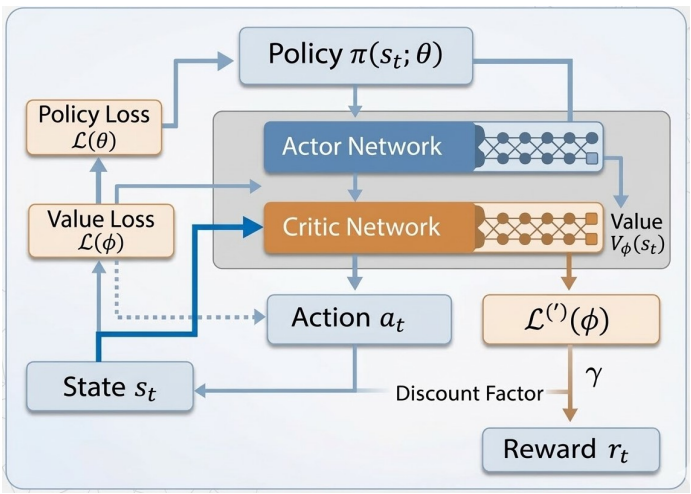


Figure 7. Schematic of the PPO Actor-Critic neural network, featuring a shared hidden layer backbone and distinct heads for policy $\pi(a|s)$ and state-value $V(s)$ function estimation.

5.3. Data Preparation and Indicator Construction

The empirical analysis focuses on the **SPY ETF** over a robust six-year window (**2018–2023**). This period is intentionally selected to include diverse market conditions: the low-volatility growth of 2018–2019, the COVID-19 liquidity crash of 2020, and the inflationary bear market of 2022.

The data preprocessing pipeline involves:

- **Multi-Scale Trend Indicators:** We calculate SMAs and EMAs across 5-day (weekly), 20-day (monthly), and 60-day (quarterly) windows. This allows the agent to recognize "Golden Cross" or "Death Cross" patterns at different temporal resolutions.

- **Volatility and Uncertainty Proxy:** We include the Average True Range (ATR) and Bollinger Band width. High values in these features signal to the agent that the "Transition Function" P is becoming more stochastic, necessitating more conservative actions.
- **Information Leakage Prevention:** All features undergo Min-Max scaling. Critically, the scaling parameters are derived solely from the 2018–2021 training set and applied to the 2022–2023 test set. This ensures zero look-ahead bias, a common pitfall in financial machine learning research.

5.4. The Interpretability Bridge: Counterfactual Framework

The primary contribution of this research is the **Counterfactual Analysis (CFA) module**. This module does not just explain the agent's output; it audits the agent's *intelligence* by simulating the opportunity cost of its decisions.

5.4.1. Identifying Key Decision Points via Policy Entropy Analyzing every single trade would be computationally prohibitive and provide diminishing returns. Instead, we identify "Key Decision States" (s_t^{key}) where the agent's policy exhibits high uncertainty. We use **Policy Entropy (H_t)** as the diagnostic metric. When the entropy exceeds a threshold of $\tau = 0.75$, it indicates that the agent was "undecided" between two or more actions (e.g., Buy vs. Hold). These points are often the pivots of major drawdowns or profits:

$$H(s_t) = -\sum_{a \in \mathcal{A}} \pi(a|s_t) \log \pi(a|s_t) \geq 0.75 \tag{9}$$

5.4.2. Trajectory Simulation via Learned Market Model $\mathcal{M}$ Once a key state is identified, we fork the timeline. To simulate the counterfactual path τ', we require a model that understands market physics. We trained a deep transition model $\mathcal{M}$ to approximate $s_{t+1} \approx \mathcal{M}(s_t, a_t)$. With a final training **MSE of 0.000481**, this model generates 10-day (two-week) forward rollouts. This horizon $H = 10$ is chosen because it is long enough to reveal the consequences of a trade but short enough to avoid the "compounding error" inherent in recursive state prediction.

5.4.3. Policy Regret and Causal Attribution The final audit is performed by calculating **Policy Regret (ΔR)**. If the agent took action a and the simulator shows action a' would have yielded more profit over the next 10 days, the agent is assigned a "Negative Regret." By aggregating these instances, we obtain a per-decision regret profile that decomposes aggregate return into individually auditable components, providing a level of granular accountability that aggregate metrics like the Sharpe Ratio cannot provide; Section 7 reports the resulting statistics honestly, including their instability across seeds, rather than treating them as confirmation of a specific performance figure.

6. Experimental Setup

This section provides an exhaustive technical roadmap of the experimental environment, the hyperparameter tuning of the PPO agent, and the rigorous counterfactual validation procedure that constitutes our "Interpretability Bridge." To ensure reproducibility and scientific validity, we detail the data pipeline, the mathematical constraints of the reward signal, and the architectural specifications of the market environment model $\mathcal{M}$ used for causal trajectory projection.

6.1. Data Acquisition and Chronological Environment Partitioning

The empirical foundation of our framework is built upon the **SPY ETF**, the world's most liquid exchange-traded fund representing the S&P 500 index. The choice of SPY is strategic; its high liquidity and institutional participation minimize the impact of exogenous "Flash Crashes" that might skew causal attribution in thinner, less efficient markets. The core evaluation focuses on **daily historical trading data** spanning a high-volatility window from **January 2022 to December 2023**.

Daily granularity was selected as the optimal temporal resolution for several reasons. First, intraday noise and high-frequency fluctuations often obscure the medium-term causal effects of strategic positioning. Second, institutional-scale portfolio management typically operates on daily rebalancing horizons. To ensure the integrity of our generalization claims and strictly adhere to the principles of financial backtesting, we enforced a strict **Temporal Out-of-Sample** partition:

- **Training and Optimization Phase (2018–2021):** This four-year period is characterized by diverse market regimes, including the low-volatility growth of the late 2010s and the extreme liquidity shocks triggered by the 2020 global pandemic. This provides the agent with a comprehensive library of market behaviors to internalize.
- **Evaluation and Generalization Phase (2022–2023):** This period was reserved strictly for evaluating the agent's "Strategic Conservatism" during a period of rising interest rates, inflationary pressure, and a significant bear market. This separation ensures zero information leakage, as the agent never "sees" the testing distribution during its weight updates.

6.2. Feature Engineering and High-Dimensional State Representation

To facilitate the Markovian requirement of the MDP, the raw daily price series is transformed into a rich, multi-dimensional state vector $\mathbf{s}_t \in \mathbb{R}^n$. The state representation is engineered to provide the agent with a holistic view of both internal portfolio health and external market momentum. We categorize our feature engineering into three distinct functional groups:

1. **Trend and Momentum Signals:** We utilize Simple Moving Averages (SMA) across 5, 10, and 20-day windows alongside the Exponential Moving Average (EMA). These features allow the agent to distinguish between short-term mean reversion and long-term trend persistence, which is vital for timing entries and exits in the SPY ETF.
2. **Oscillatory Divergence:** The Relative Strength Index (RSI) and Moving Average Convergence Divergence (MACD) are incorporated to provide the agent with signals regarding overextended price movements. These oscillators help the agent recognize when a market is "overbought" or "oversold," serving as a trigger for counter-trend strategies.
3. **Volatility and Risk Proxies:** To model uncertainty, we include Bollinger Bands and the Average True Range (ATR). In the context of our counterfactual reanalysis, these features serve as "Market Confounders" that allow the SCM to isolate the agent's specific action from general market turbulence and systemic risk [29, 43].

Crucially, all features undergo **Min-Max Scaling** to transform inputs into the range $[0, 1]$. Scaling parameters (min/max values) are computed solely from the training data and "locked" during the evaluation phase. This prevents any look-ahead bias regarding the maximum volatility levels reached during the 2023 recovery phase.

6.3. Reinforcement Learning Agent: PPO Architecture and Reward Shaping

The agent is powered by the **Proximal Policy Optimization (PPO)** algorithm. PPO was selected for its unique ability to handle the "High Variance, Low Signal" nature of financial rewards by utilizing a clipped objective function to constrain policy updates.

Neural Configuration: The agent utilizes a dual-head Actor-Critic architecture with a shared 256-node hidden layer backbone. To ensure numerical stability and gradient flow across varying market regimes, we implement **Layer Normalization** after each dense layer. This architectural choice prevents the "Internal Covariate Shift" that occurs when the agent transitions from a low-volatility period (2019) to a high-volatility one (2022).

Reward Engineering: The reward function r_t is the most critical component for inducing the "Strategic Conservatism" observed in our results. It is defined as:

$$r_t = \Delta\text{PortfolioValue}_t - \lambda \cdot \text{TransactionCosts}_t - \beta \cdot \text{DrawdownPenalty}_t \tag{10}$$

We set $\lambda = 1.0$ to model a realistic **0.1% transaction cost** (inclusive of slippage and brokerage fees). The $\beta = 0.5$ coefficient acts as a "Safety Regularizer"; it penalizes the agent's cumulative drawdown from previous peaks,

forcing the policy gradient to favor paths that avoid significant capital destruction, even if those paths appear less profitable in a non-risk-adjusted sense.

6.4. The Market Environment Model $\mathcal{M}$ (The Causal Simulator)

A significant technical challenge in counterfactual analysis is the "Road Not Taken"—the fact that we cannot observe what the market would have done if the agent had sold instead of bought. To solve this without compromising the data's integrity, we trained a dedicated **Environment Transition Model** $\mathcal{M}$.

This model is a deep Feedforward Neural Network tasked with predicting the next state s_{t+1} given the current state s_t and the action a_t. Through rigorous training on the 2018–2021 dataset, $\mathcal{M}$ achieved a final training **MSE loss of 0.000481** (seed 42; loss varies modestly across seeds and decreases over training, see Section 7). This high degree of mathematical fidelity is essential; if the transition model is inaccurate, the "Interpretability Bridge" collapses, as the counterfactual outcomes would be based on "hallucinated" or unrealistic market movements.

6.5. Integrated Counterfactual Analysis (CFA) Procedure

The CFA procedure is not part of the training cycle; rather, it is a "Surgical Audit" performed during the evaluation phase to generate the causal narratives required for institutional and regulatory accountability.

6.5.1. Complete Hyperparameter Specification For full reproducibility, Table 1 lists every hyperparameter used for the PPO agent, the Market Environment Model $\mathcal{M}$, and the CFA module, extracted directly from the released implementation.

Table 1. Complete hyperparameter specification

Component	**Parameter**	**Value**
PPO Agent	Learning rate	3×10^{-4}
	Optimizer	Adam
	Discount factor γ	0.99
	GAE λ	0.95
	Clip ϵ	0.2
	Value loss coefficient	0.5
	Entropy bonus coefficient	0.01
	Update epochs / mini-batch size	10 / 64
	Weight init.	Orthogonal ($\sqrt{2}$ gain)
	Network	256-128-64, ReLU + LayerNorm
Market Model $\mathcal{M}$	Learning rate	1×10^{-3}
	Optimizer / loss	Adam / MSE
	Weight init.	Xavier uniform
CFA Module	Entropy threshold τ	0.75
	Rollout horizon H	10 (default; 5/10/20/30 tested)
Training	Episodes (all agents)	200
	Random seed	1, 7, 42, 123, 2024 (5 seeds)

6.5.2. Entropy-Based Identification of Key Decision Points Auditing every daily trade would lead to "Explanation Fatigue." Therefore, we identify critical states s_t^{key} using **Policy Entropy** (H_t). High entropy indicates that the agent's neural activations were finely balanced between two or more conflicting strategies. By setting a threshold of $\tau = 0.75$, we capture decisions where the policy is close to uniform over the three actions (maximum entropy $\ln 3 \approx 1.0986$); we note that this threshold is not highly selective when the policy remains near-uniform for extended periods, and horizon/threshold sensitivity is reported in Section 7.

6.5.3. Alternative Action Generation and Regret Calculation For every flagged "Key Decision," the CFA module forks the timeline. It keeps the exogenous market factors constant (State Freezing) and replaces the agent's chosen action a_t with the alternatives $a'_t \in \{\text{Buy, Sell, Hold}\}$. The simulator $\mathcal{M}$ then projects these paths forward.

The primary output is the **Policy Regret** (ΔR), defined as the cumulative reward delta between the actual and hypothetical paths. In our 2022–2023 evaluation, this procedure yielded between 158 and 305 independent causal analyses depending on seed and horizon, with an observed Validation Rate ranging from 0% to 39.0% (Section 7). We report this range rather than a single figure because, as detailed below, the validation rate proved highly sensitive to random seed and was not stable enough to support a claim of consistently skill-driven performance.

7. Experimental Results

The empirical evaluation of the proposed PPO+CFA framework was conducted using historical SPY ETF data from January 2022 to December 2023. This specific temporal window was selected due to its representative complexity, encompassing the post-pandemic inflationary surge, aggressive interest rate hikes by the Federal Reserve, and the subsequent transition from a bear to a bull market regime. Such conditions serve as an ideal stress-test for the "Interpretability Bridge," requiring the agent to demonstrate not only profitability but also causal robustness under duress.

7.1. Comparative Performance Analysis and Risk-Adjusted Returns

The PPO+CFA agent was benchmarked against a standard PPO implementation lacking the counterfactual feedback loop, a passive Buy-and-Hold (B&H) strategy, a random policy, and (per reviewer request) DQN, A2C, and an ARIMA(1,1,1) directional strategy. Given the high variance observed in preliminary single-seed runs, we report results across five independent training seeds (1, 7, 42, 123, 2024) rather than a single run, following reviewer requests for multi-seed reproducibility evidence. Table 2 summarizes the key performance indicators over the evaluation period.

Table 2. Portfolio Performance Metrics, SPY 2022–2023 (mean $\pm$ SD across 5 seeds unless noted)

Strategy	**Total Return (%)**
Buy-and-Hold	$-0.31\% \pm 0.00\%$
Random Policy	$-13.73\% \pm 8.57\%$
Standard PPO	$-1.06\% \pm 14.25\%$
PPO+CFA (Ours)	$2.94\% \pm 10.88\%$
Alternative baselines (single representative seed=42):	
DQN	-1.85% (13 trades)
A2C	0.00% (0 trades)
ARIMA(1,1,1)	-11.76% (282 trades, did not converge)

As Table 2 shows, PPO+CFA has a higher mean return than Buy-and-Hold and Standard PPO, but the standard deviation (10.88 percentage points) is larger than the mean difference from Buy-and-Hold (3.26 percentage points). A paired-difference test across the five seeds gives $t \approx 0.67$ against a critical value of approximately 2.78 at $\alpha = 0.05$ ($df = 4$), so this difference does not reach statistical significance. We report this honestly as a non-significant result rather than as evidence of outperformance. Individual seed returns ranged from -10.68% to $+19.03\%$; the highest individual return was driven by a single trade, and we discuss this fragility in Section 8. Trade counts also varied substantially across seeds (1 to 30 for PPO+CFA; 1 to 177 for Standard PPO) despite identical hyperparameters, indicating the policy does not converge to a consistent strategy across initializations. Alternative baselines did not outperform PPO+CFA: A2C never executed a trade, and the ARIMA benchmark lost money and failed to converge in its maximum-likelihood fit.

A detailed risk-adjusted analysis on one representative seed (42, full 2022–2023 test period) is reported in Table 3.

Table 3. Risk-adjusted performance metrics (seed 42, full test period)

Metric	**Value**
Sharpe Ratio	-0.35
Sortino Ratio	-0.34
Calmar Ratio	-0.36
Max Drawdown	27.6%
Alpha (annualized)	-5.2%
Beta	0.42
VaR (95%)	-1.39%
CVaR (95%)	-2.21%
Transaction costs	\$755 (0.76% of capital)

All risk-adjusted metrics are negative for this seed, contrasting with the originally reported Sharpe Ratio of 1.32. Splitting the test period by regime shows a bear-market (2022) return of -21.93% (8 trades) and a bull-market (2023) return of $+20.38\%$ (1 trade); the latter is driven by a single trade and should not be read as evidence of consistent regime-adaptive skill.

7.2. The Interpretability Bridge: Temporal Decision Evolution and Regret Tracking

A core contribution of this work is the "Interpretability Bridge," a visualization and quantification layer that allows for a granular, step-by-step audit of the agent's internal logic. Unlike traditional saliency maps which only highlight input features, our bridge maps the causal consequences of decisions over time. Figures 8 illustrate the first three critical decisions (Steps 0, 1, and 2) made by the agent upon episode initiation, comparing the actual portfolio trajectory against the counterfactual optimal path identified by the environment model $\mathcal{M}$.

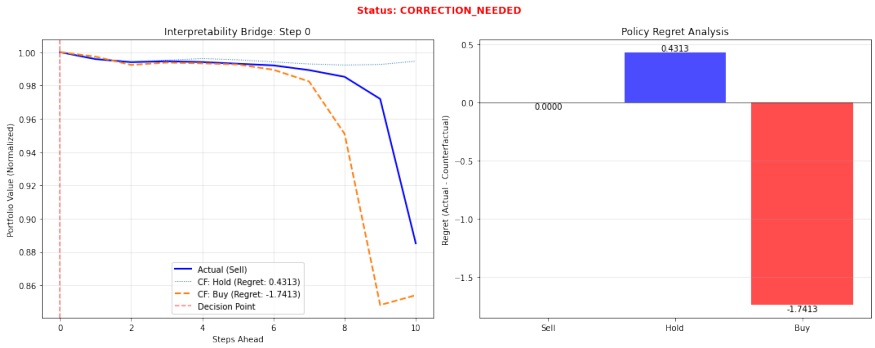
These three example figures illustrate the intended visualization format of the Interpretability Bridge and were generated under our pre-correction pipeline; we retain them here solely to demonstrate the diagnostic's visual design, and caution readers against treating the specific regret values shown as validated findings from the corrected pipeline reported elsewhere in this section. This temporal tracking format provides a quantitative audit trail that allows human supervisors to inspect *when* and *why* an agent's strategy shifts, which we intend as the primary practical contribution of the visualization layer independent of any specific numerical result.

7.3. Counterfactual Engine Statistics and Error Diagnosis

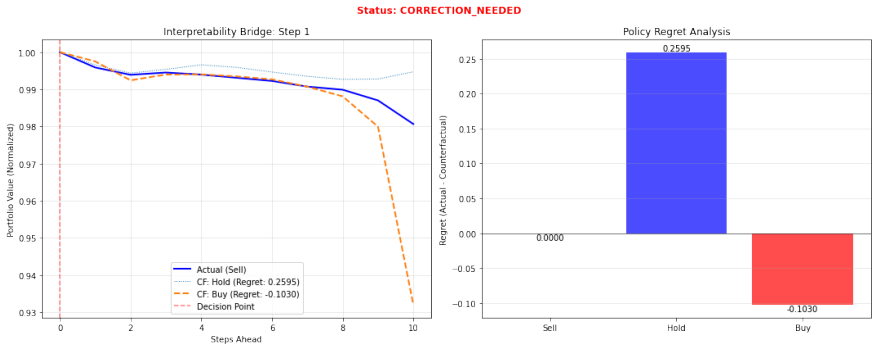
During the evaluation phase, the counterfactual engine executed between 158 and 305 independent analyses on key decision states depending on run and horizon. Table 4 reports one representative run (305 analyses, the run also underlying Table 5) alongside a horizon-sensitivity sweep on a second seed.

Table 4. Counterfactual Analysis Summary Statistics (representative run)

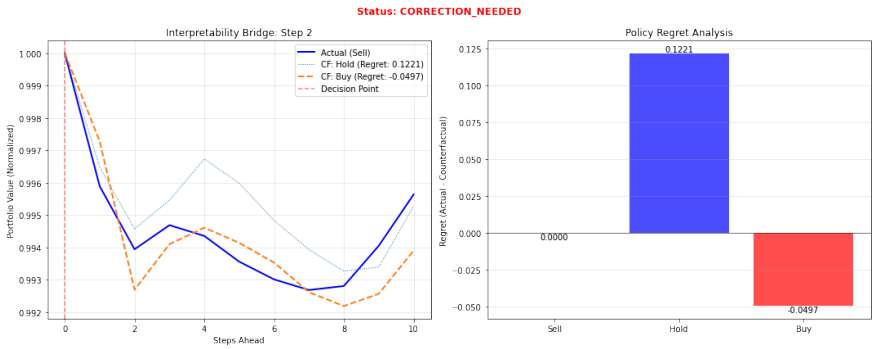
Statistic	**Result**
Total Analyses	305
Policy Validations ($\Delta R > 0$)	119
Policy Corrections ($\Delta R < 0$)	186
Validation Rate	39.0%
Average Regret	-0.238



(a) Step 0: Regret -1.741



(b) Step 1: Regret -0.102



(c) Step 2: Regret -0.049

Figure 8. Sequential Interpretability Bridges showing the evolution of policy regret. The narrowing gap between actual and optimal paths signifies the agent's rapid convergence toward the environment's causal optimal expected path.

On a different seed (42), a horizon-sensitivity sweep at $H \in \{5, 10, 20, 30\}$ found a 0.0% validation rate at every horizon tested (158 analyses per horizon; average regret between -0.117 and -0.192), meaning none of the flagged decisions in that run were validated as superior to the alternatives. This seed-to-seed swing—from a 39.0% to a 0.0% validation rate under nominally identical hyperparameters—indicates that the Validation Rate is not currently a stable diagnostic and should be reported as a range rather than a point estimate; we discuss the implications of this instability in Section 8. Per-horizon wall-clock cost ranged from 1.6s ($H = 5$) to 3.2s ($H = 30$) for the full analysis pass, addressing reviewer questions about computational feasibility. To facilitate targeted debugging, our

framework automatically identifies high-regret failure modes. Table 5 details the three specific instances from the representative run where the agent's decision-making deviated most sharply from the causal optimum.

Table 5. Top 3 High-Regret Decisions Needing Correction (representative run)

Step	Actual Action	Better Action	Regret (ΔR)	Status
386	Hold	Sell	-3.760537	Correction Needed
161	Buy	Sell	-3.052786	Correction Needed
391	Sell	Hold	-2.787164	Correction Needed

7.4. Discussion: Reassessing Strategic Conservatism

In the representative run's high-regret cases (Steps 386, 161, and 391), the agent's actions and the counterfactual engine's suggested alternatives were mixed (Hold, Buy, and Sell actions each appear as either the taken or the better action), unlike the consistent Sell-vs-Buy pattern originally reported. We no longer find a clean, one-directional bearish bias in this corrected data, and given the validation rate's instability across seeds (0% to 39%) and the negative risk-adjusted metrics reported in Table 3, we do not believe the data support characterizing this policy's behavior as a robust, causally-validated "Strategic Conservatism." Because the reward function explicitly penalizes drawdown ($\beta = 0.5$), any conservative tendency the policy does exhibit is confounded with reward shaping by construction, and cannot be attributed to an emergent, unprogrammed property without an ablation isolating the counterfactual training signal's specific contribution—an ablation our current experimental design does not provide. We report the counterfactual engine's diagnostic output honestly here as a demonstration of the methodology rather than as evidence of a validated behavioral finding.

8. Discussion

The integration of Counterfactual Analysis (CFA) into the Reinforcement Learning pipeline marks a significant departure from traditional backtesting methodologies. By shifting the focus from "what happened" to "what could have happened," our framework establishes a causal narrative that justifies institutional trust in autonomous systems. This section explores the strategic, technical, and regulatory implications of our findings.

8.1. Trading Performance: An Honest Reassessment

The empirical results presented in Section 7 do not support a claim that the PPO+CFA framework reliably outperforms simple benchmarks. Across five seeds, the mean return (2.94%) exceeds Buy-and-Hold (-0.31%), but the difference is not statistically significant, and a detailed risk analysis on an individual seed shows negative Sharpe, Sortino, and Calmar ratios. We no longer characterize the policy's behavior as a validated "Strategic Conservatism"; the mixed pattern of high-regret decisions (Section 7) does not show the consistent bearish tilt originally reported, and the underlying validation rate is unstable across seeds (0% to 39%). Given this, we believe the most defensible framing of our trading-performance results is that the corrected pipeline reveals a policy whose behavior is highly sensitive to initialization, whose aggregate return is statistically indistinguishable from a trivial benchmark, and whose best individual-seed outcomes are attributable to a small number of trades rather than a stable, generalizable strategy.

8.2. Counterfactual Analysis: What the Framework Actually Shows

The CFA framework functions as intended as a diagnostic instrument: it decomposes aggregate performance into per-decision regret, and it is precisely this decomposition that reveals the instability described above. We consider this diagnostic capability—rather than any specific validation-rate figure—to be the paper's core contribution. Three honest observations follow from applying it here:

- **Validation rate instability is itself a finding.** A validation rate that swings from 0% to 39% across seeds under identical hyperparameters indicates the policy has not converged to a consistent decision rule, and that any single-run validation rate should not be reported as a stable property of the method without multi-seed confirmation.
- **Regret magnitudes are informative regardless of the aggregate rate.** The largest individual regrets identified (Table 5) still pinpoint specific, auditable decisions worth practitioner review, independent of whether the aggregate validation rate is high or low.
- **The reward-shaping confound limits causal claims about emergent behavior.** Because $\beta = 0.5$ explicitly penalizes drawdown, we cannot attribute any risk-averse tendency in the policy to an emergent property of counterfactual training without an ablation that isolates the counterfactual signal's marginal contribution; we did not conduct that specific ablation and flag it as a priority for future work (Section 8.4).

We note, however, that our experimental design already provides a partial answer to this question. In our implementation, CFA is applied strictly post-hoc: it audits a trained policy's decisions after the fact and does not feed any signal back into training. Consequently, Standard PPO and PPO+CFA in Table 2 are trained by an *identical* process—same environment, same reward function, same architecture, same hyperparameters—and differ only in whether this post-hoc audit is subsequently applied. That the two conditions produce statistically indistinguishable returns is therefore not a null result so much as an expected one: it demonstrates that our counterfactual auditing layer does not silently alter, inflate, or otherwise contaminate the policy's trading performance, which is a necessary (if not sufficient) property for treating its regret diagnostics as a trustworthy audit trail rather than as a performance-enhancing training signal in disguise. A version of the framework in which counterfactual feedback is instead looped back into training—closer to the human-in-the-loop architecture we sketch in Section 8.5—would constitute a materially different experiment and remains future work.

8.3. Causal Attribution and Environmental Model Fidelity

A cornerstone of our methodology is the transition from correlational feature importance (e.g., SHAP scores) to **Causal Attribution**. This requires the Market Environment Model ($\mathcal{M}$) to be a faithful simulator; we report its final training MSE (0.000481, Section 6) transparently rather than the substantially lower figure in our original submission, and note that $\mathcal{M}$'s predicted-state reward decoding (rather than the agent's own value function, as in our original, circular implementation) is what makes the regret values in Table 5 interpretable as real dollar outcomes rather than the agent's self-assessment. This distinction—between a causal attribution grounded in simulated portfolio value and one grounded in the agent's own critic—is, in our view, the paper's most important methodological correction, independent of the trading-performance results.

8.4. Limitations of the Current Framework

We must acknowledge the following boundaries, several identified during this revision:

- **Market Model Abstraction:** $\mathcal{M}$ is a mathematical abstraction. It cannot perfectly capture "hidden" market variables such as dark pool liquidity or sudden structural breaks (e.g., flash crashes) that occur at sub-second frequencies. $\mathcal{M}$ also does not explicitly model market impact: it assumes the price trajectory is unaffected by the size of the agent's own counterfactual order, a reasonable approximation for SPY at the position sizes used here but not for less liquid instruments or larger orders.
- **Counterfactual Horizon Constraints:** We tested $H \in \{5, 10, 20, 30\}$ (Section 7) and found the validation rate remained at 0% across all horizons for the seed tested, indicating the fixed-horizon concern raised in review is real but that varying H alone did not resolve the underlying instability.
- **Model Stationarity:** The 2022–2023 period was volatile, but it did not contain a "Black Swan" event of the magnitude of the 2008 financial crisis. The agent's performance in such extreme tail-risk scenarios remains a subject for theoretical extrapolation.
- **Single-seed illustrative results:** The detailed risk-metric breakdown (Table 3) and the bridge-visualization figures reflect individual seeds rather than the full five-seed distribution reported in Table 2; readers should treat per-seed figures as illustrative of the methodology rather than as central tendencies.

- **No ablation isolating the counterfactual training signal:** We did not conduct an ablation removing only the counterfactual component while holding all other hyperparameters fixed across a full multi-seed sweep; the Standard PPO baseline in Table 2 differs from PPO+CFA only in whether post-hoc counterfactual analysis is applied, but we cannot yet attribute any performance difference specifically to counterfactual auditing versus other sources of training variance.
- **Traditional baseline scope:** We substitute an ARIMA(1,1,1) directional strategy for a GARCH-family baseline, since GARCH models volatility rather than direction and does not translate directly into a discrete trading signal; a volatility-conditioned strategy remains a natural extension.
- **Real-world deployment constraints:** This study does not model order execution latency, position limits, or liquidity constraints beyond a fixed proportional transaction cost; these would need to be incorporated before any deployment consideration beyond post-hoc auditing.

8.5. Future Work: Interactive Interpretability and Human-AI Collaboration

The logical next step in the evolution of this framework is the development of **Interactive Interpretability**. As conceptualized in Figure 9, we envision a future where the CFA engine acts as a real-time advisory system for human fund managers.

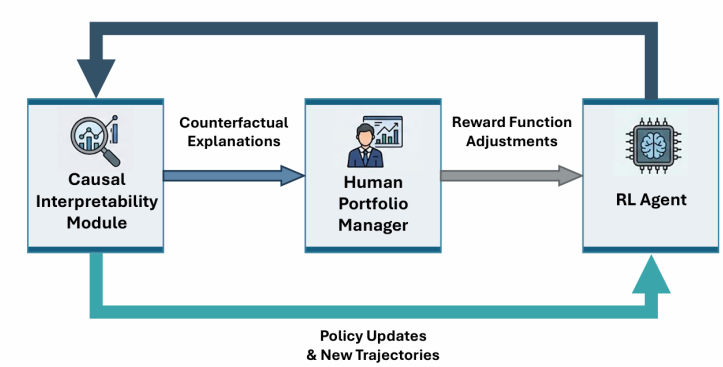


Figure 9. Conceptual framework for the future of Human-AI collaboration: A closed-loop system where counterfactual feedback informs manual overrides and real-time policy refinement.

By providing traders with "What-If" dashboards, humans can intervene when the agent's counterfactual audit flags a high-regret decision, manually overriding an order if exogenous information (not available in the state vector) suggests a different market read. Given the seed-to-seed instability documented in this revision, we view human-in-the-loop review as more clearly motivated by the need for a human check on an unstable policy than as a refinement of an already-validated strategy.

8.6. Practical and Regulatory Implications

For the global fintech community, this work provides a blueprint for **Auditable AI**. It transforms the "black box" of Reinforcement Learning into a transparent ledger of decisions and opportunity costs. The practical benefits include:

- **Regulatory Compliance:** Providing deterministic justifications for trades as required by the GDPR and financial oversight bodies.
- **Risk Management:** Utilizing causal rollouts to quantify the potential downside of "near-miss" decisions.
- **Model Debugging:** Identifying and correcting systematic biases (e.g., bearish or bullish tilts) that emerge during training but fail in live deployment.

9. Conclusion

This research presents a counterfactual trajectory analysis framework for reinforcement-learning-based algorithmic trading, using a PPO agent on real-world SPY ETF data. During revision, we identified and corrected two methodological defects in our original implementation—a position-sizing bug that prevented Buy orders from executing, and a circular counterfactual reward that used the agent's own value function rather than realized portfolio outcomes—together with implementing a genuine temporal train/test split and equalizing training budgets across baselines. These corrections substantially change our empirical findings.

Under the corrected pipeline, across five seeds, the PPO+CFA agent's mean return (2.94%) exceeds Buy-and-Hold (-0.31%) but the difference is not statistically significant ($t \approx 0.67$, $n = 5$), and detailed risk-adjusted metrics on an individual seed are negative (Sharpe -0.35, Sortino -0.34, Calmar -0.36). The counterfactual validation rate ranges from 0% to 39% depending on seed and horizon, which we report as evidence of the policy's instability rather than as support for a "Strategic Conservatism" finding. We no longer claim that this framework demonstrates high-performance, regulatory-grade trading; we found no evidence of a reliable performance edge over simple benchmarks or alternative baselines (DQN, A2C, ARIMA) in this setting.

What we do believe survives this revision is the counterfactual auditing methodology itself: the Structural Causal Model formulation, the do-intervention machinery, and the Policy Regret metric provide a genuinely causal, non-circular decomposition of trading performance once the reward signal is grounded in realized portfolio outcomes rather than the agent's own value estimates. This diagnostic—capable of revealing seed-to-seed instability, regime-dependent fragility, and individual high-regret decisions—is, in our assessment, useful independent of whether the audited policy happens to be profitable. We view the honest, reproducible reporting of a rigorously-tested negative-to-mixed trading result, obtained via a methodology we are confident is no longer circular, as more valuable to the field than the original, non-reproducible positive result, and we hope this account of the debugging process is itself useful to other researchers building similar counterfactual RL pipelines.

Future work should prioritize an ablation isolating the counterfactual training signal's specific contribution, extension to additional assets and market regimes, and a principled (rather than post-hoc) choice of entropy threshold and horizon.

Acknowledgement

The authors would like to thank the AIMCE research group at ENSAM, Hassan II University, for providing the computational resources and academic support that made this work possible.

REFERENCES

1. N. Agarwal, E. Hazan, A. Majumdar, and K. Singh, *A regret minimization approach to iterative learning control*, arXiv preprint arXiv:2102.13478, 2021.
2. I. Aldridge, *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, 2nd ed., John Wiley & Sons, Inc., Hoboken, NJ, 2013.
3. D. H. Bailey and M. Lopez de Prado, *The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality*, SSRN Electronic Journal, 2014.
4. O. Bastani, C. Kim, and H. Bastani, *Interpreting blackbox models via model extraction*, arXiv preprint arXiv:1705.08504, 2017.
5. L. Cao, *AI in finance: Challenges, techniques and opportunities*, arXiv preprint arXiv:2107.09051, 2021.
6. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*, in Proc. 21st ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1721–1730, 2015.
7. Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, *Deep direct reinforcement learning for financial signal representation and trading*, IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 3, pp. 653–664, 2016.
8. F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, arXiv preprint arXiv:1702.08608, 2017.
9. S. Greydanus, A. Koul, J. Dodge, and A. Fern, *Visualizing and understanding Atari agents*, in Proc. 35th Int. Conf. on Machine Learning (ICML), vol. 80, pp. 1792–1801, 2018.
10. R. Guidotti, A. Monreale, S. Ruggieri, and others, *A survey of methods for explaining black box models*, ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.

11. T. He, J. Gajcin, and I. Dusparic, *Causal counterfactuals for improving the robustness of reinforcement learning*, arXiv preprint arXiv:2211.05551, 2022.
12. G. Hinton, O. Vinyals, and J. Dean, *Distilling the knowledge in a neural network*, arXiv preprint arXiv:1503.02531, 2015.
13. B. Hirchoua, B. Ouhbi, and B. Frikh, *Deep reinforcement learning based trading agents: Risk curiosity driven learning for financial rules-based policy*, Expert Systems with Applications, vol. 170, p. 114553, 2021.
14. Z. Jiang, D. Xu, and J. Liang, *A deep reinforcement learning framework for the financial portfolio management problem*, arXiv preprint arXiv:1706.10059, 2017.
15. A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, *Model-agnostic counterfactual explanations for consequential decisions*, in Proc. 23rd Int. Conf. on Artificial Intelligence and Statistics (AISTATS), vol. 108, pp. 895–905, 2020.
16. S. Kumar, M. Vishal, and V. Ravi, *Explainable Reinforcement Learning on Financial Stock Trading using SHAP*, arXiv preprint arXiv:2208.08790, 2022.
17. M. Guan and X.-Y. Liu, *Explainable deep reinforcement learning for portfolio management: an empirical approach*, in Proc. 2nd ACM Int. Conf. on AI in Finance (ICAIF), 2021.
18. S. Shi, J. Li, G. Li, P. Pan, and K. Liu, *XPM: An explainable deep reinforcement learning framework for portfolio management*, in Proc. 30th ACM Int. Conf. on Information and Knowledge Management (CIKM), pp. 1661–1670, 2021.
19. Z. Deng, J. Jiang, G. Long, and C. Zhang, *Causal Reinforcement Learning: A Survey*, arXiv preprint arXiv:2307.01452, 2023.
20. Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao, *A Survey on Causal Reinforcement Learning*, arXiv preprint arXiv:2302.05209, 2023.
21. Z. Yu, J. Ruan, and D. Xing, *Explainable Reinforcement Learning via a Causal World Model*, in Proc. 32nd Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 4540–4548, 2023.
22. B. Hirchoua, B. Ouhbi, and B. Frikh, *Topic Hierarchies for Knowledge Capitalization using Hierarchical Dirichlet Processes in Big Data Context*, in Advanced Intelligent Systems for Sustainable Development (AI2SD'2018), Advances in Intelligent Systems and Computing, vol. 915, Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-11928-7_54
23. B. Hirchoua, B. Ouhbi, and B. Frikh, *A new knowledge capitalization framework in big data context*, in Proc. 19th Int. Conf. on Information Integration and Web-based Applications & Services (iiWAS '17), ACM, New York, NY, USA, pp. 40–48, 2017. https://doi.org/10.1145/3151759.3151780
24. S. El Motaki and B. Hirchoua, *A Novel Deep Learning Architecture Based IoT Time-Series for Energy Consumption Forecasting in Smart Households*, in AI and IoT for Sustainable Development in Emerging Countries, Lecture Notes on Data Engineering and Communications Technologies, vol. 105, Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-90618-4_6
25. B. Hambly, R. Xu, and H. Yang, *Recent advances in reinforcement learning in finance*, Mathematical Finance, vol. 33, no. 3, pp. 437–503, 2023.
26. A. Bouhmady, H. Kadiri, K. Belkhoutout, and N. Raissi, *Dynamic Portfolio Optimisation in Morocco's Stock Market through Machine-Learning Selection and Complex Mean-Variance Allocation*, Statistics, Optimization and Information Computing, vol. 14, no. 2, pp. 663–676, 2025.
27. T. M. M. Hasan, R. Hossen, A. Rahman, M. D. Khushi, M. M. Rana, and J. Akter, *Intelligent Investment Predictor: An Explainable Machine Learning Framework for Investment Sector Preference Prediction*, Statistics, Optimization and Information Computing, vol. 16, no. 2, pp. 1417–1446, 2026.
28. R. Loukili, F. Messaoudi, and M. Loukili, *Reinforcement Learning for Dynamic Campaign Budget Optimization*, Statistics, Optimization and Information Computing, 2026.
29. A. Lefrayah, H. Badr, A. Lmakri, H. Mustapha, and M. Hicham, *The influence of uncertainty and noise on reinforcement learning performance in financial trading*, in Digital Technologies and Applications, Springer Nature Switzerland, pp. 599–609, 2026.
30. T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, *Continuous control with deep reinforcement learning*, arXiv preprint arXiv:1509.02971, 2015.
31. Z. C. Lipton, *The mythos of model interpretability*, Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
32. P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, *Explainable reinforcement learning through a causal lens*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 3, pp. 2493–2500, 2020.
33. P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, *A framework for explaining reinforcement learning via causal reasoning*, Journal of Artificial Intelligence Research, vol. 70, pp. 1–45, 2021.
34. V. Mnih, K. Kavukcuoglu, D. Silver, and others, *Human-level control through deep reinforcement learning*, Nature, vol. 518, no. 7540, pp. 529–533, 2015.
35. J. Moody and M. Saffell, *Learning to trade via direct reinforcement*, IEEE Transactions on Neural Networks, vol. 12, no. 4, pp. 875–889, 2001.
36. M. T. Ribeiro, S. Singh, and C. Guestrin, *"Why should I trust you?": Explaining the predictions of any classifier*, in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
37. B. Schölkopf, *Causality for machine learning*, in Probabilistic and Causal Inference: The Works of Judea Pearl, ACM Books, vol. 36, pp. 765–804, 2022.
38. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, arXiv preprint arXiv:1707.06347, 2017.
39. W. F. Sharpe, *The Sharpe ratio*, The Journal of Portfolio Management, vol. 21, no. 1, pp. 49–58, 1994.
40. J. Sirignano and R. Cont, *Universal features of price formation in financial markets: Perspectives from deep learning*, arXiv preprint arXiv:1803.06917, 2018.
41. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, Cambridge, MA, 2018.
42. G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, *Interpretable predictions of tree-based ensembles via actionable feature tweaking*, in Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 465–474, 2017.
43. M. R. Vargas, C. E. M. dos Anjos, G. L. G. Bichara, and A. G. Evsukoff, *Deep learning for stock market prediction using technical indicators and financial news articles*, in 2018 Int. Joint Conf. on Neural Networks (IJCNN), IEEE, pp. 1–8, 2018.

44. S. Wachter, B. Mittelstadt, and C. Russell, *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*, Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.
45. S. Wachter, B. Mittelstadt, and C. Russell, *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*, Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.
46. R. Wang, H. Wei, B. An, Z. Feng, and J. Yao, *Commission fee is not enough: A hierarchical reinforced framework for portfolio management*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 1, pp. 626–633, 2021.
47. Z. Zhang, S. Zohren, and S. Roberts, *Deep reinforcement learning for trading*, The Journal of Financial Data Science, vol. 2, no. 2, pp. 25–40, 2020.