ADVANCING INVESTMENT FRONTIERS: INDUSTRY-GRADE DEEP REINFORCEMENT LEARNING FOR PORTFOLIO OPTIMIZATION

Philip Ndikum **Serge Ndikum**

February 2024

ABSTRACT

This research paper delves into the application of Deep Reinforcement Learning (DRL) in asset-class agnostic portfolio optimization, integrating industry-grade methodologies with quantitative finance. At the heart of this integration is our robust framework that not only merges advanced DRL algorithms with modern computational techniques but also emphasizes stringent statistical analysis, software engineering and regulatory compliance. To the best of our knowledge, this is the first study integrating financial Reinforcement Learning with sim-to-real methodologies from robotics and mathematical physics, thus enriching our frameworks and arguments with this unique perspective. Our research culminates with the introduction of AlphaOptimizerNet, a proprietary Reinforcement Learning agent (and corresponding library). Developed from a synthesis of state-of-the-art (SOTA) literature and our unique interdisciplinary methodology, AlphaOptimizerNet demonstrates encouraging risk-return optimization across various asset classes with realistic constraints. These preliminary results underscore the practical efficacy of our frameworks. As the finance sector increasingly gravitates towards advanced algorithmic solutions, our study bridges theoretical advancements with real-world applicability, offering a template for ensuring safety and robust standards in this technologically driven future.

***K*eywords** Portfolio Optimization, Artificial Intelligence, Mathematical Finance, Deep Reinforcement Learning, Robotics, Sim-to-Real Transfer, Risk Management, Regulation, Software Engineering System Design.

1 Introduction and Broader Impact

The integration of Reinforcement Learning (RL) in asset-class agnostic portfolio optimization marks the beginning of a transformative era in quantitative finance, demanding interdisciplinary expertise and novel approaches. This study exemplifies this transformative shift, introducing an industry-grade RL framework integrated with state-of-the-art (SOTA) computational techniques, all tailored for real-world financial applications. To our knowledge, this is the first research paper to adapt simulation-to-real (sim-to-real) transfer perspectives from robotics and mathematical physics to finance. Our frameworks are designed to navigate high-stakes, non-stationary market dynamics within stringent regulatory frameworks. Arguably, this multidisciplinary approach enhances the practicality and robustness of financial models, paving the way for future research that rigorously accounts for historical risks [1]–[10]. The domain of Deep Reinforcement Learning (DRL) has seen significant breakthroughs across various complex fields, underlining its potential to tackle previously insurmountable challenges. Among these, the success of DeepMind's AlphaGo Zero in mastering the ancient game of Go stands as a testament to RL's capability in navigating intricate strategic landscapes. Reinforcement Learning's effectiveness in complex games, particularly in non-stationary environments, with partial information and multi-agent dynamics, closely mirrors the intricate challenges faced in financial markets. Recent advancements in RL, especially in diplomatic simulation games, underscore its potential in navigating stochastic decision-making processes under uncertainty and incomplete information. Such developments are emblematic of RL's adaptability and skill, mirroring the complexities of financial portfolio management and signifying the technology's

applicability in this sophisticated domain [11]–[16]. In robotics, leading companies like Boston Dynamics, DiDi, Tesla and others are leveraging Reinforcement Learning to develop advanced autonomous robots and vehicles. This highlights RL's versatility and effectiveness in various high-stakes scenarios [17]–[20]. In a parallel development, Large Language Models (LLMs) have undergone significant evolution through Reinforcement Learning with Human Feedback (RLHF). These models have not only achieved progress in natural language processing but also show promise in specialized domains like finance. Preliminary research suggests that financial LLMs may eventually outperform traditional human analysts in analyzing and interpreting complex financial data, indicating a potential shift in how financial information is processed and utilized [21]–[24].

Our research endeavors to bridge the gap between theoretical concepts in academia and the dynamic, often unpredictable realities of real-world financial markets. Despite commendable and ambitious goals, many financial RL papers and open-source software artifacts, such as FinRL and its derivatives, often do not fully meet the stringent requirements of robust software engineering, regulatory compliance, and practical market constraints [25]–[28]. We hope that this paper offers insights to enhance industry-grade applications of Reinforcement Learning in portfolio optimization and Financial RL more broadly. Our collaboration was born from our observation that many available open-source tools lack consistent requirements, robust documentation, and often exhibit coding errors stemming from inadequate finance domain knowledge. It is worth noting, that the arguments presented also align with industrial trends in the Artificial Intelligence (AI) industry, with many practitioners advocating for improved engineering and software design practices [29]–[34]. In economics and AI research, replication challenges and incomplete findings, have been addressed in many excellent papers by industry practitioners but few have provided frameworks to concretely address these challenges in modern finance [35]–[43]. Given the high stakes of financial management, as evidenced by past financial crises and the fluctuating performance of investors, it is crucial that RL algorithms deployed in finance undergo rigorous stress-testing, statistical experiment design and conform to high engineering standards [44]–[48]. The proprietary library and RL agent, *AlphaOptimizerNet*, were meticulously crafted to ensure robustness in experiment design and performance analysis[2]. Embracing an iterative and system-driven design philosophy, the development acknowledges the nuanced complexities inherent in contemporary AI system design. This research underscores the potential significance of sim-to-real considerations, a central theme throughout this paper. Such an approach may be crucial for organizations operating in an ecosystem increasingly shaped by advanced algorithms. Our objective is to argue for the adoption of elevated standards in both academic and industrial realms, contributing to the future landscape of financial management. The insights presented in this research might be particularly relevant to sovereign wealth funds, family offices, pension funds and public policy makers. In today's financial environment, these institutions play a pivotal role as guardians of capital for the benefit of families, citizens, and nations. We aim to enhance individual well-being and positively influence the broader trajectory towards global prosperity.

2 Transcending Traditional Portfolio Optimization: A Reinforcement Learning Approach

2.1 Modern Portfolio Theory: Limitations and RL's Role in Future Directions

Portfolio optimization is a fundamental component of quantitative investing, tasked with the strategic allocation of assets to maximize returns while minimizing risk. It encompasses a broad spectrum of asset classes and investment timelines, reflecting the multifaceted nature of financial markets. In our research, we consider portfolio optimization as a prime example to demonstrate the application of our Deep Reinforcement Learning (RL) methodologies. Building on the work of Ndikum (2020) [1], we highlight the critical role of financial, econometric and regulatory domain knowledge for the successful deployment of advanced AI algorithms in finance. As we delve into Deep RL techniques for portfolio optimization, we consciously choose to employ simplified mathematical formulations. This approach, aimed at making our work accessible to a diverse audience from various fields, reflects our commitment to balance technical detail with general comprehensibility. Through this strategy, we aim not only to engage a broader readership but also to foster meaningful discourse across different disciplines, maintaining the essence and integrity of our arguments throughout [3]. Our objectives in this section are twofold: Firstly, we aim to elucidate the fundamental concept of diversification and the mechanics underlying Modern Portfolio Theory (MPT). Secondly, we delve into the inherent limitations of MPT, particularly in the context of today's dynamic financial markets. A thorough grasp of these foundational concepts is indispensable, as it lays the groundwork for a deeper appreciation of the sim-to-real framework presented later in our paper. In the realm of portfolio management, diversification is not merely a strategy for risk mitigation;

[2]It is important to note that many developments in the fields of finance and investment are proprietary and thus not publicly disclosed. As a result, this paper solely references publicly available academic literature and does not speculate nor comment on the advancements or strategies employed within private institutions. The development of our RL agent, *AlphaOptimizerNet*, and the accompanying proprietary library, is an example of such non-public, specialized work in the field.

[3]For a comprehensive textbook on Reinforcement Learning, we recommend *Reinforcement Learning and Stochastic Optimization: A unified framework for sequential decisions* by Professor Warren Powell, of Princeton University, 2022. [50].

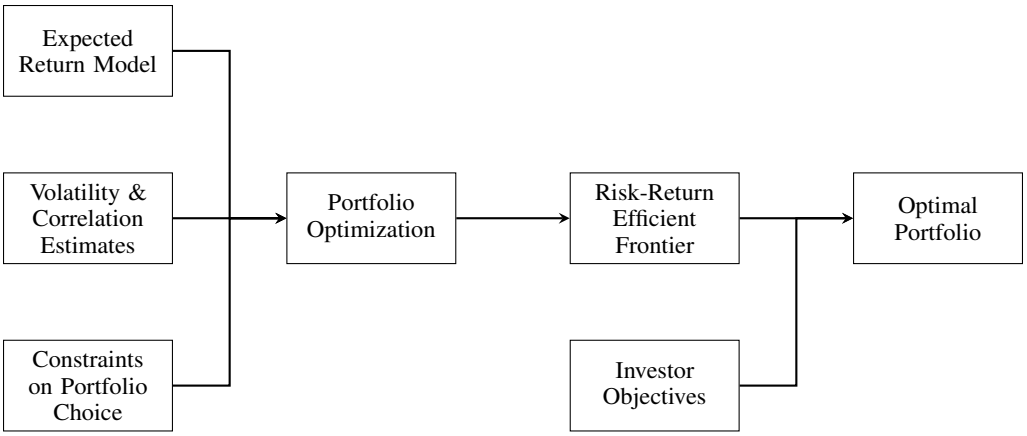

Figure 1: This diagram, adapted from Fabozzi et al. (2002) [49], succinctly outlines the Modern Portfolio Theory (MPT) investing process. It shows the progression from expected return and volatility & correlation modeling, through the inclusion of portfolio constraints, to optimization, culminating in the establishment of the risk-return efficient frontier and selection of the optimal portfolio.

instead it fundamentally forms the philosophical underpinning of Modern Portfolio Theory (MPT). At the time of its inception, MPT represented a paradigm shift in financial portfolio management, systematically harnessing the principles of diversification to mathematically balance risk and return. This innovative approach revolutionized the way portfolios were constructed, moving beyond intuitive strategies to a quantifiable and objective models for optimization. Diversification, a central tenet of Modern Portfolio Theory (MPT), can be quantitatively expressed by considering the expected risk, represented as the variance of returns σ_p^2, in a diversified portfolio of n distinct assets. This approach to risk quantification considers the expected fluctuations in portfolio returns. We follow a mathematical formulation similar to Bodie et al. (2020) [51]:

$$\sigma_p^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \text{Cov}(r_i, r_j). \tag{1}$$

Equation 1 calculates the expected variance of portfolio returns, σ_p^2, illustrating how diversification impacts this risk[4]. Here, w_i and w_j represent the portfolio weights allocated to assets i and j respectively, and $\text{Cov}(r_i, r_j)$ is the covariance between their returns, measuring how their returns move in relation to each other. The essence of diversification in risk management lies in combining assets with varying degrees of correlation, thereby reducing the portfolio's overall expected risk. The influence of this diversification becomes more significant as the number of assets increases and their correlations diversify. This dynamic is captured in the following equation:

$$\lim_{\substack{n\to\infty,\\ \rho_{ij}\to 0}} \sigma_p^2 = \lim_{\substack{n\to\infty,\\ \rho_{ij}\to 0}} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \text{Cov}(r_i, r_j) = C = \inf_{i\in\mathcal{N}} \sigma_i^2, \tag{2}$$

Equation 2 captures the theoretical impact of diversification on the expected risk of a portfolio, measured using the portfolio variance σ_p^2. As the number of assets n in the portfolio grows indefinitely and their pairwise correlations ρ_{ij} converge to zero, the portfolio's expected risk trends towards a theoretical minimum limit, denoted as C. This limit, C, is specifically defined as the infimum ($\inf$) of the variances (σ_i^2) of the individual assets within the portfolio set $\mathcal{N}$. In essence, C represents the lowest level of risk (variance) that is theoretically achievable under the ideal scenario of infinite diversification with completely uncorrelated assets. This concept is central to portfolio theory, emphasizing the advantage of having a diverse mix of uncorrelated or negatively correlated assets in a portfolio. Such diversification is a key strategy in risk management, aiming to reduce the impact of market volatility and improve the overall stability of an investment portfolio. In practical terms, it means spreading investments across a variety of asset classes to buffer against unexpected market movements, thereby enhancing the resilience of one's financial portfolio. The use of infimum notation in this equation highlights the inherently positive nature of this theoretical minimum risk. It underscores that even with ideal diversification, portfolio risk doesn't reduce to zero but converges towards the lowest variance found among the individual assets.

[4]In this context, the notation σ_p^2 represents the variance of the portfolio's returns, and ρ_{ij} (Greek Rho) refers to the correlation coefficient between the returns of assets i and j.

This concept of minimizing expected risk through diversification leads us to the core principles of Modern Portfolio Theory (MPT), as formulated by Harry Markowitz. In recognition of their pivotal contribution to the field, Markowitz, along with Merton Miller and William Sharpe, were collectively awarded the Nobel Prize in Economics in 1990 [52]–[55]. MPT pragmatically applies the principle of diversification in the context of portfolio optimization. MPT algorithms are often categorized under Quadratic Programming (QP) or Mixed Integer Programming (MIP) problems within the fields of numerical analysis and applied mathematics. Definition 2.1 provides a mathematical formulation of a single-period MPT, drawing upon the formulations by Chang et al. (2002) and Cesarone et al. (2013) [56]–[60]. In this definition, n represents the total number of financial assets, w_i represents the portfolio allocation to asset i in the range $[0, 1]$. The expected return and covariance for each asset are denoted as μ_i and σ_{ij}, respectively. In simple terms, we aim to optimize portfolio weights to minimize risk while targeting a specific expected return $\mathbb{E}[R] = R$. This unconstrained formulation can be solved efficiently with modern computational tools, allowing for the calculation of optimal portfolio weights, designated as $\phi(\mathbb{E}[R])$. The resultant *efficient frontier* identifies portfolios that optimize expected return for various levels of risk within the interval $[R_{\text{min}}, R_{\text{max}}]$. In real-world investing scenarios, this deceptively simple formulation of Modern Portfolio Theory (MPT) is significantly complicated by the introduction of various real-world constraints.

Definition 2.1 (Markowitz Single Period Unconstrained Portfolio Optimization[56], [57])**.**

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} w_i w_j \\ \textit{s.t.} \quad & \sum_{i=1}^{n} w_i \mu_i = \mathbb{E}[R] \\ & \sum_{i=1}^{n} w_i = 1 \\ & w_i \in [0, 1], \textit{ for } i \in \{1, \ldots, n\}. \end{aligned} \tag{3}$$

The Financial literature published in the last few decades has thoroughly explored and addressed these constraints, adding complexity to MPT models. The intricate and high-stakes nature of finance necessitates a deep dive into this rich history. While contemporary and future technological advancements in Reinforcement Learning, present innovative possibilities, their effective application demands a thorough grounding in the enduring realities of financial markets and regulatory environments. Neglecting the depth of historical financial knowledge can oversimplify intricate market dynamics, potentially steering strategies towards catastrophic outcomes for financial institutions and investors in both the private and public sectors. Below, we enumerate some constraints and place them in context with real-world problems:

1. **Cardinality Constraint:** This constraint limits the number of assets in a portfolio to a maximum of n:

$$|\text{supp}(w)|_0 \leq n, \quad \text{where } \text{supp}(w) := \{i : w_i > 0\}. \tag{4}$$

   Here, $\text{supp}(w)$ denotes the set of indices i for which the investment in asset i (w_i) is greater than zero, implying active inclusion of the asset in the portfolio. The expression $|\text{supp}(w)|_0$ counts the total number of distinct assets with non-zero investments in the portfolio. The constraint $|\text{supp}(w)|_0 \leq n$ ensures that this number does not exceed the predetermined maximum n, thereby limiting the total number of different assets included in the portfolio. In asset management, this constraint is relevant for ensuring portfolio transparency and manageability.

2. **Volume Constraint:** The volume constraint restricts the allocation to each asset in the portfolio within a specified range, denoted by

$$\ell_i \leq w_i \leq u_i \text{ for each asset } i \in \{1, \ldots, n\}, \tag{5}$$

   where ℓ_i and u_i define the lower and upper bounds, respectively, for the investment in asset i. This constraint is critical in preventing over-concentration in a single asset, thereby reducing the risk of significant loss from adverse movements in that asset's value. In high frequency trading (HFT), the volume constraint indirectly influences turnover and liquidity, both of which are critical factors affecting the overall effectiveness of trading strategies.

3. **Regulatory Constraints and Transaction Costs:** Legal and regulatory constraints may dictate investment limitations in certain sectors or geographies. For example, contemporary investment strategies may be

influenced by client or government mandates that encourage investments in certain sectors, such as renewable energy, reflecting broader objectives that may not be captured mathematically by classical models. These external directives can significantly shape portfolio composition. Beyond regulatory considerations, transaction costs, including taxes and brokerage fees, play a crucial role. The incorporation of such constraints can dramatically alter the performance of back-tested simulations in both Modern Portfolio Theory and newer methodologies such as Reinforcement Learning.

As we can observe, the adoption of Modern Portfolio Theory (MPT) in dynamic financial scenarios necessitates confronting a myriad of complexities. Additionally, we note that these classical solutions operate under static market assumptions and are often geared towards single-period frameworks. This inherent limitation impedes their utility in environments characterized by asymmetric risk profiles[5], human irrationality, and complex game-theoretic dynamics. It is within this context that the concept of multi-period optimization solutions have gained recent prominence:

Definition 2.2 (Multi-period Optimization)**.** *Consider a universe of n assets over a time horizon of K periods. We provide a similar formulation to Lezmi et al. (2022) [63] and transition to a probabilistic framework to model our multi-period optimization as follows:*

$$W^* = \arg\max_{W \in \Omega} \mathbb{E}\left[U(W) \mid \mathcal{F}_t\right] \tag{6}$$

where

$$W = \begin{pmatrix} w_{1,t+1} & w_{1,t+2} & \cdots & w_{1,t+K} \\ w_{2,t+1} & w_{2,t+2} & \cdots & w_{2,t+K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,t+1} & w_{n,t+2} & \cdots & w_{n,t+K} \end{pmatrix},$$

denotes the matrix of asset allocations across K future time periods, with each $w_{i,t+j}$ being the column vector of portfolio weights allocated to asset i at time period $t+j$. Here, W^* represents the optimal series of allocations. The function $U(W)$ represents the inter-temporal utility function, expressing the investor's preference over varied portfolio allocations. The term $\mathcal{F}_t$ is the filtration, representing the cumulative information available up to time t, including past market data and events. The optimization is subject to the condition $W \in \Omega$, where Ω denotes the sample space of all feasible sequences of portfolio allocations, subject to a set of linear and non-linear constraints. This framework allows for the ranking of allocations based on desirability, with superior utility values indicating more favored allocations, factoring in considerations such as returns, risks, and other investor inclinations.

The multi-period optimization shown in Definition 2.2 aims to forecast and optimize portfolio weights across several forward-looking timeframes. Despite its theoretical robustness, real-world application of multi-period optimization presents substantial challenges adequately addressed and summarized by Kolm et al. (2014) in their analysis of portfolio optimization's evolution and challenges:

> *"In practice, multi-period models are seldom used. There are several practical reasons for that. First, it is often very difficult to accurately estimate return and risk for multiple periods, let alone for a single period. Second, multi-period models are in general computationally intensive, especially if the universe of assets considered is large. Third, the most common existing multi-period models do not handle real-world constraints . . . For these reasons, practitioners typically use single-period models to rebalance the portfolio from one period to another"* - 60 Years of portfolio optimization: Practical challenges and current trends, Kolm et al. (2014) [64].

This complexity is compounded in practical numerical solutions to MPT, such as Mean-Variance Optimization (MVO). MVO applies the principles of MPT in a quantitative framework to identify the optimal asset allocation for a given risk level. A key requirement in MVO is for the covariance matrix, which represents the relationships between the returns of different assets, to be positive semi-definite. This ensures that the calculated variance of the portfolio, a key risk measure in MPT, is always non-negative. Additionally, the inherent variability or volatility in financial markets adds another layer of complexity. The assumption of stable volatility, a cornerstone of MVO, often does not hold true in real-world scenarios where market conditions can rapidly change. In contrast, Reinforcement Learning (RL) offers a promising alternative. RL thrives in environments characterized by uncertainty and the need for sequential decision-making, aligning well with the dynamic nature of financial markets. Unlike MVO, RL does not require the stringent assumptions about volatility and matrix properties. RL's ability to handle complex, dynamic environments, coupled with its flexibility in integrating real-world constraints, makes it a formidable tool in addressing the challenges highlighted by Kolm et al. [65]–[68]. State-of-the-art RL algorithms offer the potential to navigate the evolving

[5]Post-Modern Portfolio Theory (PMPT) seeks to mitigate challenges associated with asymmetric risk. Nevertheless, it predominantly utilizes conventional optimization methods, which are subject to analogous computational challenges [61], [62].

landscape of risks and large datasets, holding the promise of superior and autonomous risk-adjusted returns. This shift in focus towards RL and similar advanced algorithms, however, must not eclipse the importance of grounding these technologies in the contemporary and historical realities of financial markets and regulatory environments [69], [70]. Given the stakes involved in managing client capital, it seems imperative for the financial industry, similar to other high-risk sectors where RL is currently being used in industry (e.g., robotics), to establish rigorous standards ensuring the responsible deployment of these algorithms. In this paper, we advocate for a paradigm where the effectiveness of techniques such as RL is gauged not merely by their theoretical prowess, but by their alignment with the dynamic - and highly regulated demands of contemporary international finance which we will explore in the subsequent sections.

2.2 Reinforcement Learning in Portfolio Optimization: Bridging Theory and Practice

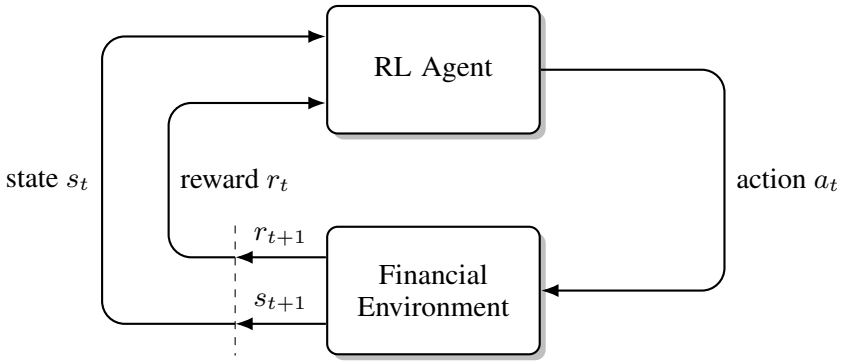

Figure 2: A schematic representation of the interaction between the RL Agent and the financial environment in the context of portfolio optimization. The RL Agent represents the algorithm making portfolio decisions, where the state (s_t) denotes the current market conditions and portfolio configuration, the action (a_t) corresponds to portfolio adjustment decisions, and the reward (r_t) reflects the financial outcome, such as risk-adjusted returns, of these decisions. This diagram illustrates how RL adapts and responds to evolving financial scenarios, thereby optimizing portfolio performance.

The transition to Reinforcement Learning (RL) in financial portfolio management marks a pivotal shift from conventional strategies to advanced, algorithm-driven approaches. Deep Reinforcement Learning (DRL), an extension of RL, effectively employs deep neural networks to navigate complex, high-dimensional environments. This subject will be further explored in upcoming sections. RL, set apart from the single-period, deterministic models of Modern Portfolio Theory (MPT), adopts a dynamic, data-driven approach. It is adept at managing multiple, even *infinite time horizons* signaling a move towards creating computational equivalents of human traders and investment managers. Reinforcement Learning systems operate as an autonomous agents (advanced algorithm or models), learning and adapting continuously via interactions with a simulated environment that mirrors real-world market complexities. This learning process, entirely driven by data, enables the RL agent to identify patterns and make informed decisions based on extensive market data. This continual, dynamic learning process is essential for developing investment strategies that are responsive to the rapidly changing and unpredictable financial markets. This evolution underscores a departure from the traditional reliance on human emotional resilience in financial decision-making. With its capacity to process large data volumes, RL holds the potential to match or even exceed human capabilities in identifying and capitalizing on market inefficiencies. Free from human psychological biases and fatigue, RL systems introduce a major transformation in the financial sector. They combine human expertise with the precision of computational systems. In this new era, trust in financial decision-making extends beyond human judgment to the dependability and accuracy of these meticulously engineered computational systems. This transition represents a major shift in the landscape of financial portfolio management. To facilitate a deeper understanding and bridge the gap towards our novel sim-to-real frameworks, we employ analogies from robotics throughout our discussion. These analogies not only provide a familiar reference point but also align closely with the principles and challenges inherent in both fields:

- **State** (s_t): In RL, the state at time t, s_t, represents a snapshot of the current situation, analogous to the perception of an environment by an RL agent in robotics. In portfolio optimization, s_t encapsulates market conditions, including key financial indicators such as asset prices and trading volumes. Understanding the state's formulation requires a nuanced grasp of the asset class and regulatory context, comparable to the detailed environmental awareness necessary for an RL agent in robotics. The strategic design of the state space, similar to sensor range and resolution considerations in robotics, is crucial for capturing detailed market representation while maintaining computational manageability. This process is central to optimizing the RL

model's efficiency in interpreting and reacting to market dynamics, paralleling the way an RL agent in robotics processes sensory data to understand its environment.

- **Action** (a_t): The action a_t at time t signifies the decision made by the RL agent, comparable to decision-making processes of RL agents in robotics. In portfolio optimization, this action primarily involves determining the asset allocations, represented as a vector. The constraint $\sum_{i=1}^{n} a_i = 1$ ensures that the total allocation across all assets equals one, reflecting a similar allocation of resources or efforts by RL agents in robotics. This constraint is critical in ensuring that the portfolio weights are proportionally distributed. Actions may also incorporate factors such as maintaining cash reserves or opting for periods of inactivity, dependent on the strategy. The action space design must align with specific investment strategies and comply with regulation and investment mandates, including decisions regarding the use of leverage.
- **Reward** (r_t): The reward r_t at time t functions as the guiding incentive for the RL agent, comparable to the objectives pursued in robotics where agents learn to navigate and interact with dynamic and complex physical environments. In portfolio optimization, the agent's reward is typically aligned with the maximization of risk-adjusted returns, tailored to meet investor-specific goals and limitations. This mechanism is crucial in guiding the agent's learning path, enabling it to develop and refine strategies that effectively balance return and risk.
- **Environment:** In financial RL, the environment represents the complexities of market dynamics, regulatory frameworks, and economic indicators, similar to the intricate and evolving environments encountered in robotics. The financial environment is inherently non-stationary and only partially observable, posing significant challenges comparable to those faced in robotic navigation and interaction. Designing an RL model for finance requires a deep understanding of market intricacies and nuances. The environment should mirror specific investment timeframes and strategies, with considerations varying greatly between applications such as high-frequency trading and long-term asset management. It's imperative that the environment is realistically modeled to encapsulate market behaviors and constraints, ensuring the RL model's relevance and effectiveness in actual financial scenarios.

As we explore Reinforcement Learning (RL) for portfolio optimization, understanding its mathematical foundations is crucial. RL algorithms, though complex, rely on well-established mathematical principles and computational procedures. Creating these systems necessitates expertise in AI, finance, and a deep understanding of regulatory and risk considerations. This skill set is essential for successful deployment of robust Reinforcement Learning systems in finance. Our aim is to present these mathematical concepts clearly and comprehensively, particularly for readers with a finance background. While equations are our focus, the accompanying explanations provide an insight into their underlying principles. Designing RL systems for portfolio optimization requires more than technical proficiency; it entails integrating financial knowledge, AI capabilities, and an understanding of real-world regulations and risks. As we delve into Markov Decision Processes (MDPs), Partially Observable Markov Decision Processes (POMDPs), and key aspects of Deep Learning, our goal is to maintain a balance between technical rigor and accessibility. Subsequent sections will cover mathematical equations while contextualizing them within financial portfolio management.

2.3 Mathematical Formalism: MDPs, POMDPs, Deep Learning

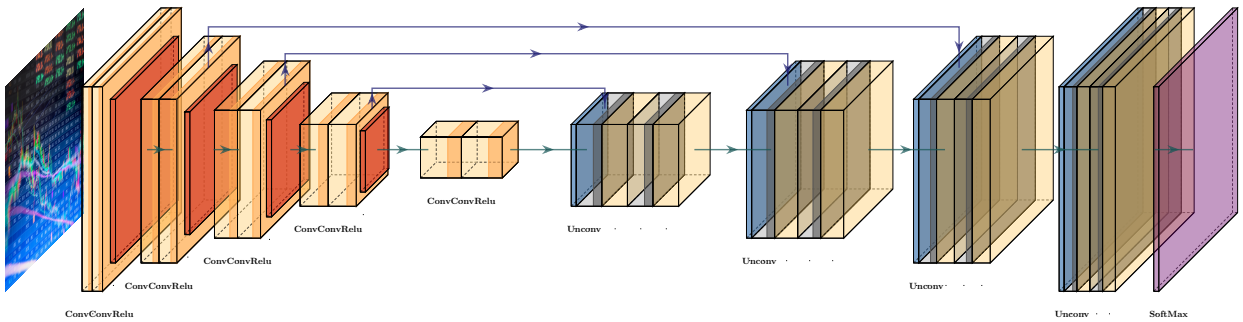

Figure 3: This figure displays a U-Net, originally a Convolutional Neural Network for biomedical imaging, now adapted to exploit patterns in financial data. Its interoperability underlines the crucial role of interdisciplinary expertise in crafting advanced RL systems for finance.

We now delve into the mathematical formalism of Deep Reinforcement Learning (RL) shown in Fig. 2. This section is structured to first introduce intuitive explanations of Markov Decision Processes (MDPs) and Partially

Observable Markov Decision Processes (POMDPs), followed by their formal definitions, and then a discussion of Deep Learning and neural networks. A key distinction between broader Reinforcement Learning (RL) approaches and Deep Reinforcement Learning (DRL) is the use of deep neural networks in the latter, enabling them to handle more complex, high-dimensional and noisy financial market data. Although an in-depth technical exploration exceeds the scope of this paper, we introduce neural networks as the key driving force behind our RL algorithms. Analogous to *computational brains*[6], these biologically inspired algorithms are indispensable due to their ability as universal function approximators, essential for effective planning, prediction, and decision-making within MDP and POMDP frameworks. In our examination of Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs), let's offer simplified interpretations before presenting formal definitions. MDPs represent decision-making frameworks with complete market information. Imagine making investment decisions where you have full visibility into asset prices, economic indicators, and market trends. This scenario aligns with an MDP framework. Conversely, POMDPs apply when some market information is hidden or uncertain. Picture a situation where certain variables, such as insider information or future market movements, are unknown. Despite this uncertainty, decisions must be made based on available information. POMDPs provide a framework for decision-making under such conditions, allowing adaptation to market dynamics even with incomplete visibility:

Definition 2.3 (Markov Decision Processes in Portfolio Optimization)**.** *In portfolio optimization, an MDP is characterized by the tuple $(S, A, T, T_0, R, \gamma)$. Here, S represents the array of possible market states, encompassing variables like asset prices and economic indicators. A denotes the range of possible investment actions, reflecting decisions across a portfolio of n assets, thus $A \subseteq [0,1]^n$. The transition function $T : S \times A \to \Delta(S)$ models the probability of moving from one market state to another, given a specific investment action. $T_0 : S \to [0,1]$ defines the initial distribution of states in the market. The reward function $R : S \times A \to \mathbb{R}$ quantifies the financial impact of each action, considering both returns and associated risks. Unlike finite-horizon models, the time horizon in this setting is considered infinite, and the discount factor $\gamma \in [0,1)$ reflects the long-term strategy of the investment approach. The primary objective is to determine an optimal policy $\pi^* : S \to A$ that maximizes expected returns over this infinite horizon, as formalized in the following equation:*

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{s_t, a_t, r_t \sim T, \pi} \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid s_0 \right]. \tag{7}$$

The objective of the RL agent in this framework is to learn the optimal mapping of states to actions (portfolio allocations), thereby maximizing long-term investment returns while adapting to the dynamic nature of financial markets.

Definition 2.4 (Partially Observable Markov Decision Processes in Financial Markets)**.** *POMDPs extend the MDP framework to scenarios where full knowledge of market states is not directly observable by investors. A POMDP in finance is characterized by the tuple $(S, A, O, T, T_0, R, \gamma)$. Here, S denotes the complete range of market states, A represents the range of investment actions, and O encompasses observable market factors. The transition function $T : S \times A \to \Delta(S)$ models the dynamics of market state changes following investment actions, and $T_0 : S \to [0,1]$ defines the initial state distribution. The observation function $O : S \times A \to \Delta(O)$ determines the likelihood of observing specific market indicators given the state and action. The reward function $R : S \times A \to \mathbb{R}$ quantifies the financial impact of actions. The decision-making policy in a POMDP, $\pi : \mathcal{H} \to A$, where $\mathcal{H}$ is the history of observations, actions, and rewards, depends on both the current and past observations. This history up to time t is represented as $\tau_{0:t} = (o_0, a_0, o_1, r_1, \ldots, a_{t-1}, o_t, r_t)$. The objective in a POMDP setting is to identify an optimal policy π^* that maximizes expected returns over an infinite horizon, considering both observable and unobservable market factors:*

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{s_t, a_t, o_t, r_t \sim T, O, \pi} \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid o_0 \right]. \tag{8}$$

In financial markets, where complete state information is often unavailable, this model becomes crucial, necessitating decision-making based on partial observations [73], [74].

In our exploration of decision-making frameworks, such as MDPs and POMDPs within dynamic financial environments, we have encountered foundational concepts that underpin adaptive learning. Now, we pivot our focus to a fundamental concept in Reinforcement Learning (RL) that serves as a linchpin for adaptive decision-making: the Bellman equation. Developed by Richard Bellman in 1957 [75], the Bellman equation plays a pivotal role in dissecting decision-making processes into interconnected steps, enabling the evaluation of both immediate rewards and the anticipated utility

[6]It is essential to understand that neural networks, while biologically inspired, do not replicate the full complexity of biological brains. These statistical learning algorithms are abstractions, simplifying key principles from our evolving understanding of neuroscience. Interestingly, there appears to be a correlation between advancements in computational neuroscience and the development of increasingly performant neural network architectures. However, this should not be mistaken for a comprehensive understanding of the brain's intricacies or an implication of neural networks' complete mimicry of biological processes [71], [72].

of future actions. At the heart of RL methodologies lies the concept of the value function, which encapsulates the expected cumulative future rewards an agent can attain from a given state. In the context of portfolio optimization, this value function serves as the cornerstone of the RL agent's learning process, guiding it towards determining the optimal portfolio allocation strategy. By iteratively updating the value function based on observed rewards and transitions between states, the RL agent gradually hones its understanding of the market dynamics and learns to make informed decisions that maximize long-term returns while balancing risk. With this understanding of the value function's significance in RL, let us delve into the formulation of the Bellman equation and its implications for portfolio optimization:

$$V(s_t) = \max_{a_t} \left[R(s_t, a_t) + \gamma \sum_{t'=t+1}^{\infty} P(s_{t'}|s_t, a_t) V(s_{t'}) \right] \tag{9}$$

where:

- **Value Function** $V(s_t)$: Serves as a predictive model that evaluates the expected future performance of a portfolio from a given state s_t, representing current market conditions, asset valuations, and portfolio composition. By calculating the cumulative expected returns, the value function incorporates both immediate gains and future prospects, considering the probabilities of various market scenarios and their timings. It guides the RL agent in assessing the long-term effectiveness of different investment strategies, taking into account factors such as market volatility, portfolio diversification, and the risk-return balance. Essentially, $V(s_t)$ aids in algorithmic decision-making by projecting the future value of a portfolio under different conditions and strategies, facilitating the selection of optimal investment actions.
- **Immediate Reward** $R(s_t, a_t)$: Quantifies the instantaneous gain or loss resulting from an action a_t taken in state s_t. It reflects the immediate financial impact of trading decisions in portfolio management, such as the realized profit and loss after executing a trade. Within reinforcement learning, the concept of reward shaping is crucial, involving the design of the reward function to align with specific investment objectives and risk profiles. For instance, a hedge fund targeting absolute returns might shape $R(s_t, a_t)$ based on log returns to emphasize direct profitability. Conversely, a strategy focused on risk-adjusted returns might utilize a reward function derived from the Sharpe ratio or its variants, thereby incorporating risk considerations. Effective reward shaping plays a pivotal role in guiding the RL agent toward investment strategies that align with investor goals, striking a balance between immediate returns and risk management.
- **Discount Factor** γ: Determines how the RL agent weighs immediate versus future rewards. In finance, it functions similarly to the discounting of future cash flows to their present value. A higher value of γ indicates a stronger emphasis on long-term gains, mirroring a strategic long-term investment approach in finance. By adjusting γ, the RL model can be fine-tuned to prioritize either short-term gains or long-term investment objectives, aligning with the specific goals and risk tolerance of the investor.
- **Transition Probability** $P(s_{t+1}|s_t, a_t)$: Represents the probability of moving from the current state s_t to a new state s_{t+1} after taking action a_t. In the financial context, this is similar to the probability of market shifts resulting from specific investment actions. It captures the essence of market volatility and the inherent unpredictability in financial decision-making. This concept in RL reflects the need to account for the variable nature of markets, where each action can lead to multiple potential future scenarios, each with its own likelihood and financial implications.
- **Value of Subsequent State** $V(s_{t+1})$: Reflects the expected utility for future states resulting from current actions. In financial terms, it is comparable to projecting the performance of a portfolio into the future, taking into account the potential outcomes of present investment decisions. This aspect of RL encapsulates the idea of forward-looking analysis in portfolio management, where the focus is not only on immediate results but also on how current choices shape future financial landscapes. It highlights the importance of strategic planning and anticipatory decision-making in both RL and financial investment, considering the long-term implications of actions taken today.

Transitioning from the traditional state-value focus of the Bellman equation, shown in Equation 9, our approach in financial portfolio optimization embraces a broader perspective. In classical Reinforcement Learning, the concept of an optimal policy, denoted as π^*, is central. This policy is typically formulated to maximize the expected cumulative reward over time, as shown in the standard RL formulation:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right] \tag{10}$$

Within the financial context, our emphasis shifts to identifying an optimal policy π^* that maximizes expected utility over an infinite horizon. This shift is encapsulated in the following formulation for Financial RL, where the optimal

policy is interpreted as the *optimal algorithmic investment strategy*:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t U(r_t, \sigma_t, X_t, w_t) \right] \tag{11}$$

In this formulation, π^* is not merely focused on maximizing immediate returns but integrates a broader spectrum of financial metrics, including returns (r_t), risk (σ_t), external market factors (X_t), and portfolio weights (w_t). This multi-faceted approach reflects the intricate nature of financial decision-making, where balancing risk, return, and adapting to market dynamics are paramount:

- **Optimal Policy** π^*: Denotes the strategy that maximizes expected utility over an infinite horizon, which is crucial in financial contexts. In Financial RL, this is comparable to identifying the most effective algorithmic investment strategy, tailored to the dynamics and complexities of financial markets.
- **Expected Utility Maximization** $\mathbb{E}_{\pi}$: Represents the expectation of utility under policy π, quantifying the long-term effectiveness of an investment strategy. This concept aligns with our earlier discussion on maximizing expected outcomes in MPT, subject to real-world constraints as detailed in Equation 6.
- **Recursive Utility Function** $U(r_t, \sigma_t, X_t, w_t)$: Integrates critical financial metrics at each time step t, including returns (r_t), risk (σ_t), external market factors (X_t), and portfolio weights (w_t). This function is pivotal in reflecting the nuanced trade-offs between risk and return, aligning closely with investor objectives and their specific financial goals.

In our previous discussions, we differentiated between standard Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL), noting the integration of neural networks in the latter. Now, we delve deeper into this distinction, focusing specifically on the role of neural networks and Deep Learning architectures within DRL. These sophisticated computational frameworks are pivotal in advancing portfolio management strategies, offering enhanced predictive and adaptive capabilities far beyond the scope of traditional RL. The following discussion will delve into the complexities of these neural networks and Deep Learning architectures, emphasizing their critical role in financial portfolio optimization. Renowned for their efficacy, Deep Learning architectures form the backbone of DRL, processing complex, high-dimensional, and non-stationary financial data with unparalleled efficiency. The advent of diverse architectures such as Convolutional Neural Networks (CNNs) for image and time-series data, Recurrent Neural Networks (RNNs) for sequential data, and Transformers[7] for complex text and language processing tasks, has revolutionized how large and noisy financial datasets are handled. These architectures are adept at extracting pertinent patterns and insights from a wide array of data types, ranging from numeric and textual to audio and visual formats, significantly broadening their applicability across diverse financial contexts. This adaptability is evidenced by their demonstrated superiority over traditional econometric models when applied to asset classes and sectors such as energy, real estate, equities, and exotic derivatives [76]–[80]. Furthermore, recent scientific advancements have extended Deep Learning's capabilities in long-term forecasting, challenging prevailing industry misconceptions [81], [82]. This progress in Deep Learning, underpinned by a generalized mathematical formulation of neural networks, lays the foundation for understanding their operational framework in financial portfolio optimization:

Definition 2.5 (Generic Representation of Deep Neural Networks [81], [82])**.** *A deep neural network with L layers can be mathematically represented as a composition of L functions $f_i : E_i \times H_i \to E_{i+1}$, where E_i, H_i, and E_{i+1} are inner product spaces for each $i \in [L]$. The state variables are denoted as $x_i \in E_i$ and the parameters as $\theta_i \in H_i$. The network's output for an input $x \in E_1$ is defined by a function $F : E_1 \times (H_1 \times \cdots \times H_L) \to E_{L+1}$, expressed as:*

$$F(x; \theta) = (f_L \circ \cdots \circ f_1)(x), \tag{12}$$

where θ encapsulates the parameter set $\{\theta_1, \ldots, \theta_L\}$. This high-level formulation adapts to various architectures. Each layer function f_i boosts adaptability and learning, crucial for tasks like optimal portfolio allocations in RL. The head map α_i and tail map ω_i aid in understanding network behavior and facilitate backpropagation for neural training. In financial portfolio optimization, the final layer f_L might act as the strategic mastermind, comparable to an advanced trader, in an MDP or POMDP setting, positioning the neural network as the strategic core of these decision processes.

In summary, our exploration of Reinforcement Learning (RL) in financial portfolio management has been comprehensive. Beginning with foundational concepts such as states, actions, and rewards, we progressed to understanding the recursive nature of decision-making through the Bellman equation. This journey led us through the intricacies of Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs), demonstrating their application

[7]Transformers have gained popularity not only in time series forecasting but also in language processing tasks. They are widely used in Large Language Models (LLMs), such as those popularized in ChatGPT by OpenAI, for tasks like text generation, translation, and summarization.

in modeling dynamic financial markets. The integration of neural networks as computational engines within these frameworks was pivotal, enabling nuanced predictions essential for portfolio optimization. However, transitioning theoretical models to real-world financial markets presents challenges beyond technical complexities. It necessitates a blend of financial expertise, AI knowledge, and regulatory understanding to ensure system robustness, rigorous model validation, and compliance with regulatory and game theoretic constraints. While RL offers solutions beyond Modern Portfolio Theory (MPT), its application in finance is not without complexities and risks, demanding careful design and implementation to balance innovation with risk management. Having established a solid understanding of RL's theoretical and practical aspects in finance, we now turn to historical parallels between technology and finance. This exploration sets the stage for our novel sim-to-real frameworks, grounded in RL advancements, aiming to revolutionize portfolio management. Through empirical testing via simulations, we aim to validate our theoretical propositions and demonstrate the real-world viability of our innovative approaches.

3 Advancing Portfolio Management: Novel Frameworks in Reinforcement Learning

As we embark on presenting our novel frameworks in the realm of financial portfolio management, this section first establishes a foundational context, underscoring how modern finance has historically drawn inspiration from various scientific and technological disciplines. By delving into the intersections between technology and finance, we aim to provide a well-grounded rationale for our innovative approaches, rooted in the rich tapestry of interdisciplinary influences that have shaped financial strategies over time. Our exploration begins by highlighting the parallels between financial strategies and technological advancements, with a particular focus on recommendation systems. Both fields have navigated similar optimization challenges, employing Reinforcement Learning (RL) as a transformative tool. This convergence underscores RL's potential to significantly impact finance, a domain that has traditionally embraced methods from physics and probability theory, notably in areas such as options pricing and risk management [70], [83]. Building on this historical context, we then introduce our novel *simulation-to-real* (*sim-to-real*) frameworks. These frameworks, inspired by advancements in robotics and mathematical physics, are specifically designed to address the complexities of modern financial portfolio management. They represent a strategic and innovative response, reflecting finance's history of adapting cross-disciplinary methodologies. This is crucial given the high-stakes nature of and unique challenges faced by the financial sector. The final computational experiments aim to rigorously evaluate our *sim-to-real* frameworks, gauging their potential impacts on the industry. Positioned at the cutting edge of current Artificial Intelligence research, these frameworks aim to stimulate discussion and the development of scalable, efficient, and practical solutions in finance. Our approach, rooted in historical precedents and cross-disciplinary insights, anticipates a future where finance is increasingly influenced by advanced algorithms. In advocating for their responsible and effective application, we emphasize the need for these innovative technologies to be meticulously designed to meet the complex and highly regulated demands of modern international finance.

3.1 Parallel Evolution: Reinforcement Learning in Finance and Technology

Building on our earlier discussions of Modern Portfolio Theory (MPT) in finance, this subsection examines the striking similarities between optimization processes in finance and technology, especially when confronted with predefined constraints. In finance, Modern Portfolio Theory (MPT) aims to maximize returns by strategically allocating investments across a mix of assets such as stocks, bonds, and derivatives, balancing risk and potential gains. Each choice is subject to a myriad of real-world constraints including risk profiles, market conditions, and regulatory frameworks. In finance, the optimization challenge resembles the task encountered in technology's recommendation systems. These systems are designed to construct portfolios of digital content, encompassing a diverse range of offerings from products to movies and social media advertisements. Their objective is twofold. Firstly, to maximize expected revenue for the firm by strategically aligning offerings with user interests. Secondly, to enhance the user's experience by precisely tailoring recommendations. This user-centric approach aims to ensure that users are presented with content that resonates with their preferences, geographical locations, and personal interests. The success of these recommendation engines is measured not only by traditional engagement metrics like click-through rates and viewing time but also by the degree to which they fulfill user expectations and drive revenue growth. This meticulous strategy in content curation mirrors the principles of portfolio optimization in finance, where the goal is to balance risk and return to meet specific investment objectives. The historical development of both finance and technology reveals profound methodological and mathematical parallels. In finance, numerical optimization methods under MPT are employed to derive the best risk-adjusted returns for portfolios. This mirrors technology's early recommendation systems, which used mathematical techniques such as matrix factorization to predict user preferences for digital content [84]–[87]. Central to both domains is the concept of utility maximization. In finance, utility functions are employed to balance risk against return, aligning with investor objectives. Conversely, in technology, utility functions gauge user satisfaction, focusing on digital engagement and content preferences. Furthermore, both fields confront analogous challenges. The *cold start* problem

in finance, characterized by initiating portfolio allocations without prior data (in the case of new financial products), is comparable to the challenge in technology of recommending content to new users who lack engagement history. Additionally, the issue of *sparse rewards* is a common obstacle in both domains [88], [89]. In finance, especially in contexts such as options trading, the concept of sparse rewards is manifested in the form of delayed financial outcomes. Until the expiry of an option or its exercise, the profit and loss (PnL) remains uncertain. This delay in realizing PnL can range from minutes to days or even months, depending on the nature of the investment. Similarly, in the realm of technology, particularly in recommendation systems, sparse rewards are seen as delayed monetary returns. The financial benefit for a technology firm materializes only when a user takes a definitive action, such as clicking on an advertisement or making a purchase from recommended products. Here too, the delay before seeing PnL can vary significantly, from immediate responses to prolonged periods of user engagement. These parallels in problem-solving structures and strategies between financial portfolio management and technology-based recommendation systems underscore a fundamental mathematical symmetry. This symmetry provides a logical foundation for hypothesizing the potential for Reinforcement Learning (RL) to transform finance, drawing on its successful applications in technology. This analysis not only reinforces the connection between these domains but also highlights the potential for cross-domain insights to inform future advancements in financial portfolio optimization.

For financial professionals, this parallel offers valuable insights. The evolution and resolution of these challenges in the technology sector, particularly through the adoption of advanced Reinforcement Learning techniques, suggest a potential roadmap for similar advancements in financial portfolio management. By acknowledging and learning from the technological sector's approach to these analogous problems, finance professionals can envisage and prepare for potential revolutionary shifts in portfolio management strategies. For those within the finance sector who may harbor skepticism about the transformative potential of advanced algorithms, the evolution and successes of RL in technology present a compelling case. As advancements in technology continued to evolve, the classical methods in both portfolio optimization and recommendation systems encountered significant limitations in addressing the dynamic and complex nature of their respective domains. This is particularly evident in portfolio optimization, where conventional methodologies rooted in Modern Portfolio Theory (MPT) often face challenges adapting to the realistic constraints and high-dimensional, noisy data characteristics of modern markets. The technology sector, in its quest for more adaptive and scalable solutions, increasingly embraced Deep Learning and Reinforcement Learning. RL's capability to learn from user interactions and dynamically adapt strategies proved to be evolutionary in addressing the intricate challenges of recommendation systems. The shift from static, algebraic solutions to dynamic, data-driven RL approaches indicates a potential evolutionary path for financial portfolio management. Notably, in the technology sector, issues once deemed computationally intractable have been effectively resolved by elite teams of mathematicians and physicists at major technology firms. These groups leverage innovative RL systems, achieving high efficiency, low latency, and scalability, thus enabling real-time recommendations for millions of users [90]–[95].

This historical parallel, where two distinct fields faced similar mathematical challenges and evolved towards RL-based solutions, strongly suggests that the future of finance is poised to be reshaped by these advanced algorithmic approaches [92]–[102]. Much like how advanced algorithms revolutionized the advertising industry - a field once dominated by marketing and business experts - finance is poised to undergo a similar transformation. Sophisticated algorithmic solutions may challenge and reshape traditional practices, potentially altering the roles and tasks traditionally associated with investors. In envisioning a future increasingly driven by advanced algorithms in finance, we assert the need for prudence and sustainability in their application, especially given the stringent regulatory environment and the immense responsibility of managing client funds. Given the unique complexities of the financial sector and the responsibility of managing substantial funds, the deployment of Reinforcement Learning techniques must be conducted with meticulous care. Unlike recommendations in social media or e-commerce, errors in financial portfolio optimization can lead to risk exposures amounting to hundreds of millions or billions of dollars, with far-reaching consequences for investment firms, financial institutions, and sovereign wealth funds. Thus, the integration of RL into finance necessitates not only advanced algorithmic development but also a profound comprehension of market nuances and robust risk management protocols. Particularly in leveraged scenarios, the potential financial ramifications underscore the importance of rigorous risk assessment and stress testing practices.

Collaboration is crucial in these practices, necessitating alignment among scientific, quantitative, risk, and regulatory teams to ensure RL's responsible utilization in high-stakes environments. Furthermore, academic researchers in this field are obliged to communicate assumptions and potential risks transparently. In integrating advanced technologies like Reinforcement Learning into financial portfolio optimization, effective collaboration among engineers, scientists, finance and legal professionals is paramount. This multidisciplinary synergy is vital, not only to drive innovation but also to ensure thoughtful application of RL within the highly regulated and complex landscape of finance. In the technology sector, the popular adage *move fast and break things* encapsulates a culture that values rapid innovation and risk-taking. However, when considering the application of similar technological advancements in finance, this approach necessitates cautious reconsideration. A reckless importation of this ethos into the financial domain could lead to dire

consequences, potentially sparking severe financial crises and triggering stringent regulatory responses, which could hinder future innovation. This paper advocates for a more deliberate and sustainable approach - *move prudently and build sustainably* - as a guiding principle in the development of financial technology. Our aim is to spark and contribute to discussions about the future of financial technology, highlighting the need for a balanced approach that recognizes the unique risks and responsibilities of the financial sector. As we progress, we will delve into hypothetical case studies that offer simplified yet insightful perspectives on the unique challenges faced in finance as opposed to technology, emphasizing the importance of an interdisciplinary, cautious and collaborative approach.

Practical Implications in Financial Portfolio Management: In applying Reinforcement Learning to portfolio optimization, it's crucial for academic and industry research to transparently justify methodologies and reasoning. The diverse nature of financial markets and regulatory environments means that theoretically sound strategies might vary significantly based on specific market conditions and investor objectives. Consequently, the developers of such mathematical and computational artifacts must clearly articulate the assumptions underpinning the chosen approach and their respective rationale. This clarity is vital for understanding the real-world applicability and constraints of these sophisticated mathematical models in the complex landscape of financial portfolio management and optimization[8]. To illustrate these points, we present a few simplified open-ended questions, highlighting the significant complexity of building industry-grade RL systems in portfolio optimization and financial RL more broadly:

1. **Understanding Time Horizon & Market Microstructure:** The transferability of an RL agent from one market domain to another poses significant challenges. Take, for instance, an RL agent (algorithm) developed for portfolio optimization in the energy derivatives market. The question arises: is it appropriate or even feasible to apply this agent to a disparate context such as a cross-asset strategy? Such a shift would involve transitioning from a market with specific characteristics to one with potentially vastly different microstructures and trading dynamics. This case underscores the need for domain-specific knowledge and strategy alignment. Moreover, the time horizon of the trading strategy plays a crucial role. An RL agent trained for high-frequency trading, where decisions are made on a millisecond basis, may not be suitable for long-term, long-only investment strategies. This distinction is rooted in the unique market microstructures and investment objectives associated with different trading styles and asset classes. It exemplifies the importance of understanding both the time horizon and the microstructural nuances of the market when deploying RL agents in finance [103]–[105].

2. **Understanding Dynamical Systems & Non-Stationarity:** The adaptability of RL agents to dynamic market conditions is put to the test in scenarios such as the onset of the COVID-19 pandemic. An agent trained in pre-pandemic market conditions might face significant challenges in adjusting to the subsequent, drastically altered market dynamics. This raises critical questions: has the agent been rigorously stress-tested to handle the volatility and shifts in market behavior seen post-COVID? Are the risk metrics utilized by the agent, sufficiently robust and adaptable to suit different investment styles and client risk profiles in this new market landscape? This scenario accentuates the inherent difficulty in dealing with non-stationary environments in financial markets, a problem that even state-of-the-art RL models grapple with. It underscores the importance of a comprehensive design of experiments [44]–[48], and continual adaptation and stress testing in financial RL applications. Such a meticulous approach is crucial for developing RL systems that are not only theoretically sound but also resilient to market evolution.

3. **Understanding Asymmetric Risk & Game Theory Dynamics:** A critical aspect of applying RL in finance is navigating asymmetric risk profiles and the game-theoretic nature of financial markets. This complexity is rooted in the constantly changing rules and dynamics that govern financial systems, unlike the more stable natural laws observed in the natural sciences. In the financial world, the introduction of new regulations, sudden market shocks, or strategic moves by major market players can drastically alter the market environment. This fluidity presents unique challenges for RL agents, which must be capable of understanding and responding to asymmetric risks and the strategic behavior of other market participants. It leads to a scenario where the data available for training and validation of these agents is not only limited but also rapidly becomes outdated as market dynamics evolve. This situation renders financial RL arguably more complex than many other domains, making the development of these respective systems among the most intricate. The crucial question then becomes: are the RL agents designed to account for and adapt to these asymmetric risks and game-theoretic complexities? [106]–[108]

[8]The development of RL systems for financial applications is an intricate endeavor that encompasses detailed mathematical modeling and considerable computational efforts. It demands not only expertise in algorithmic design but also adherence to stringent software engineering standards. Minor implementation errors can lead to significant operational complications, underscoring the need for precision in both the mathematical and software engineering facets of RL. This highlights the importance of creating robust, reliable, and efficient systems for high-stakes financial applications.

This discussion sheds light on the nuanced notion of *fully autonomous* agents in finance, which, upon closer examination, appears somewhat of a misnomer. Finance, with its intricate web of risk management and regulatory compliance, necessitates a human-in-the-loop (HIL) approach, essential for overseeing and validating algorithmic decisions [109]–[112]. Taylor et al. (2023) articulate this necessity well: "*[One should] not think of RL as a fully autonomous paradigm, but instead as an iterative learning and development process involving both learning algorithms and humans. While better algorithms may chip away at this assumption, full autonomy in terms of problem identification, construction, and deployment is unlikely in the near-future. Instead, we argue that it is critical to consider how humans and RL agents can work together to solve sequential decision tasks*" [110]. In concluding this section, we have emphasized the vital role of human insight and expertise in guiding and understanding the application of Reinforcement Learning (RL) in portfolio optimization and broader financial contexts. Effective RL systems in finance must navigate a landscape filled with uncertainties that often extend beyond the scope of sophisticated statistical algorithms. The presented scenarios highlight this blend as critical for the successful and responsible application of RL in finance.

3.2 Simulation-to-Real Transfer in financial RL: Bridging the Reality Gap

Drawing upon the successful application of Reinforcement Learning in high-risk domains like mathematical physics and robotics, we introduce our novel *Simulation-to-Real* (*Sim-to-Real*) frameworks for financial portfolio optimization. This advancement is underpinned by the interdisciplinary convergence of applied mathematics, computational science, and finance. These fields collectively reveal synergies and analogous methodologies, shaping our algorithmic solutions for financial Reinforcement Learning. Central to our approach is Sim-to-Real transfer learning, a method pivotal in the development of high-stakes, autonomous systems such as self-driving cars. Here, RL systems are exhaustively tested in simulated settings that closely mimic diverse real-world scenarios. This methodical simulation is critical in preparing the systems to navigate and adapt to the multifaceted and unpredictable nature of real-world environments, effectively narrowing the *reality gap* – the often observed divergence between simulated models and real-world outcomes. In the realm of finance, the reality gap is characterized by the unique challenges posed by fluctuating market conditions, regulatory changes, and trading constraints. Our RL models for financial portfolio optimization are meticulously crafted to be resilient, adaptable, and progressively evolving in response to the dynamic nature of financial markets. We emphasize a strategic application of RL, deeply rooted in the practical realities of market dynamics and regulatory frameworks. By integrating Sim-to-Real principles from sectors like robotics and mathematical physics, our goal is to significantly enhance the practicality and flexibility of RL in financial portfolio optimization. This approach not only aligns with our broader research objectives of fusing theoretical insights with real-world market complexities but also advocates for meticulous engineering, comprehensive domain expertise, and the responsible implementation of financial RL systems. Such a strategy echoes the critical insights gained from robotics, underscoring the importance and intricacies of deploying Sim-to-Real transfer in the financial sector:

> *"There was wide agreement that the ultimate goal is to design robotic systems that live in the real-world . . . A key aspect in sim-to-real transfer is the choice of simulation. Independently of the techniques utilized for efficiently transferring knowledge to real robots, the more realistic a simulation is the better results that can be expected . . . Reinforcement learning algorithms often rely on simulated data to meet their need for vast amounts of labeled experiences. The mismatch between the simulation environments and real-world scenarios, however, requires further attention to be put to methods for sim-to-real transfer of the knowledge acquired in simulation"* - Perspectives on sim2real transfer for robotics, Höfer et al (2020) [113].

The advancements in high-risk robotics, as discussed by Höfer et al (2020), have been achieved through addressing significant technical challenges and innovations. As we apply sim-to-real constraints and methodologies similar to those in robotics to our financial RL models, we anticipate encountering comparable challenges. This underscores the need for innovation and meticulous application in finance, a domain with considerably high stakes. Our literature review on Reinforcement Learning and Artificial Intelligence in robotics has led to the formulation of three fundamental principles. These principles are not only grounded in advanced computational science but are also acutely aware of the complexities and challenges unique to financial markets. Our objective is to establish a preliminary framework tailored to the specific needs of financial portfolio management, while also adhering to the stringent regulatory frameworks that govern this sector. Drawing on insights from a wide array of fields, including robotics and mathematical physics [114]–[144], we now introduce our novel *Simulation-to-Real* (*Sim-to-Real*) frameworks for financial RL. These frameworks are distinctively designed to address the specific challenges and nuances of financial RL use-cases:

1. **Realistic Reward Shaping & Environmental Modeling:** Central to our sim-to-real transfer methods in high-risk financial applications is the precise construction of financial environments mirroring real-world market conditions. This requires in-depth modeling of transaction costs, market impact, and regulatory constraints, essential for effective RL strategies. Recent academic efforts, like using Generative Adversarial Networks

(GANs) for market data generation, highlight the challenge of accurately replicating market dynamics. Without realistic simulation, such techniques risk being ineffective. The *domain randomization* approach in robotics, where RL models face diverse simulated conditions to build resilience, is enhanced by incorporating concepts like *concentrability*, which ensures training on a representative sample of the state space, and *coverability*, which guarantees exposure to a wide range of market scenarios including rare but critical (black-swan) events. These additions are crucial for RL systems to effectively navigate and adapt to the volatile nature of financial markets. Reward shaping is equally crucial, requiring the selection of metrics such as log returns or the Sortino ratio, tailored to specific financial goals and investor profiles. This domain-specific knowledge is imperative for aligning RL agent actions with investor objectives. Furthermore, robust software engineering is key, particularly with open-source financial modeling libraries. The final RL software must be theoretically sound, resilient, and scalable, capable of operating efficiently in diverse computational environments, including industrial-grade investment settings.

2. **Robust Risk Analysis & Statistical Stress Testing:** In the diverse and complex landscape of finance, it is imperative to design Reinforcement Learning systems that are specifically tailored to the unique characteristics of different asset classes and market microstructures, rather than attempting to create universally applicable models. This nuanced approach necessitates a comprehensive risk management strategy that incorporates conventional risk metrics, fine-tuned for particular asset classes and market sectors. A critical aspect of this approach is conducting extensive statistical analyses of performance metrics across a wide range of RL training scenarios. An active area of RL research also includes advanced causality-driven model explainability techniques, such as counterfactual explanations, which can significantly enhance understanding and trust in these systems. Emphasizing the importance of rigorous software engineering, especially for those transitioning from the computer science domain to financial applications, our methodology includes the utilization of advanced tools for serialization and cloud-based storage of performance data. This practice not only facilitates easy retrieval of crucial data during regulatory audits or client interactions but also ensures a high degree of transparency and accountability. By carefully considering these factors, we aim to develop RL systems that are not only robust in theoretical simulations but also adaptable and reliable in the real-world financial ecosystem, respecting the intricacies of each asset class and market microstructure.

3. **Strategic Resilience & Fault Tolerance-Aware Engineering:** Drawing parallels from high-risk robotics, such as autonomous vehicle systems, financial RL systems require robust fault tolerance mechanisms to anticipate and effectively manage market extremities and external shocks. This entails engineering contingency protocols for scenarios where algorithms might malfunction or underperform, particularly during volatile market conditions. Leveraging strategies utilized in autonomous robotics to handle unpredictable events, financial RL systems need proactive design considerations for navigating market anomalies and systemic fluctuations. The burgeoning threat of adversarial attacks, particularly relevant for RL models processing sentiment data from fluctuating sources like social media, must be addressed. These models require rigorous vulnerability assessments to prevent potential exploitations that could adversely affect financial outcomes. A judicious approach might involve selectively excluding data sources where the risk-to-reward ratio is unfavorable. This proactive stance on system security and fault tolerance mandates exhaustive risk assessments, encompassing both traditional financial risks and those unique to algorithmic investments, such as data manipulation and model over-fitting. Integrating these comprehensive risk management strategies from the outset ensures the development of financial RL systems that are not only resilient but which also adhere to the stringent safety and adaptability standards observed in high-risk robotics applications.

In our exploration of sim-to-real transfer methodologies from high-risk domains such as robotics to financial Reinforcement Learning (RL), we have established preliminary yet foundational guidelines. These guidelines are devised to guide the development of sophisticated RL strategies in the nuanced domains of portfolio optimization and broader financial RL. While they represent a significant initial contribution, it is essential to recognize that they are just the inception of a broader journey in this evolving field. These principles, we hope, will ignite ongoing discussions and foster safe yet innovative advancements, especially resonating with stakeholders in financial institutions, governments, and policymaking bodies worldwide. As the financial sector increasingly integrates complex algorithms, comprehensively understanding and addressing the multifaceted nature of these systems is vital. Our framework, while thorough in its current form, is not exhaustive and represents an initial step in aligning theoretical RL models with the dynamic and unpredictable nature of financial markets. With a focus on environmental realism, adaptability to market changes, and resilient system design, our work is committed to developing RL methodologies that can effectively meet the diverse challenges of the financial sector. Designed for both theoretical robustness and practical applicability, these methodologies reflect our dedication to pragmatic and effective solutions in financial management. As we conclude this part of our research, we approach a crucial stage; empirical validation of our theories through computational experiments with RL agents and our proprietary systems and market environments.

4 Computational Experiments

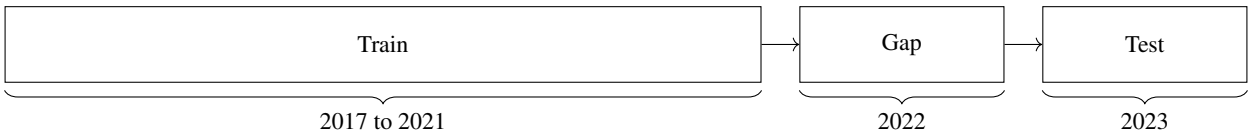

Figure 4: Experimental design setup: Our study employs Reinforcement Learning (RL) agents trained on historical data within our proprietary sim-to-real environments. In contrast to traditional backtesting approaches, our RL agents dynamically adjust portfolio weights across assets based on learned non-linear policies, offering a more adaptive and nuanced investment strategy driven by data. Evaluation is conducted over a single test year to assess the efficacy of the trained agents.

Experiment Focus & Evaluation Metrics: Our computational experiments aim to rigorously evaluate the novel *simulation-to-real* framework, as delineated in previous sections. This framework is not only centered on applying state-of-the-art (SOTA) Reinforcement Learning algorithms, recognized for their effectiveness in literature under conditions of low or negligible transaction costs, but also on a comprehensive consideration of real-world financial factors. Our experimental design integrates key real-world factors, including market frictions and broker fees, surpassing traditional academic models to provide a realistic simulation of financial market conditions. It's important to note that certain implementation details are omitted due to the proprietary nature of our algorithms and system design. This led us to adopt a simplified experimental design that emphasizes a holistic, realism-driven approach to financial market simulations. We believe that this approach, while possibly affecting algorithmic performance, provides invaluable insights for both researchers and practitioners in industry and academia. As discussed in the previous sections, we posit that from a practitioner's standpoint, an algorithm rigorously stress-tested in realistic settings holds greater value, even if it demonstrates suboptimal performance. This is in contrast to algorithms that excel in academic research but may not withstand the complexities of real-world financial environments. Our work aims to bridge this gap, offering a framework that balances theoretical robustness with practical applicability. Additionally, our experiment focus, is on specifically tailored to long-term horizon asset managers and market makers. This distinction is vital, as the strategies and constraints relevant to these entities often differ markedly from those employed by independent retail proprietary traders or entities seeking pure alpha returns. By being transparent about this focus, we ensure that our experimental design aligns closely with the operational realities and regulatory considerations unique to these segments of the financial industry. These experiments utilized high-performance cloud computing and advanced GPUs, conforming to industrial-grade software engineering practices. As part of our evaluation methodology, we incorporate a comprehensive set of performance metrics for asset-class agnostic evaluation, as suggested by Lim, et. al (2019) [145] and Zhang et. al (2020) [100]. These metrics collectively offer a comprehensive and nuanced view of portfolio performance, addressing various dimensions of risk and return, crucial for a holistic evaluation of any trading strategy:

Metric	**Definition**	**Explanation**
$\mathbb{E}[R]$	Annualized expected trade returns.	Indicates potential profitability and sets a baseline for performance expectation.
std(R)	Annualized standard deviation of trade returns.	Measures the volatility of returns, crucial for understanding the risk level of the strategy.
DD	Annualized standard deviation of negative trade returns.	Captures the strategy's downside risk, critical for gauging potential losses.
Sharpe	Sharpe ratio: $\mathbb{E}[R]/\text{std}(R)$.	Evaluates risk-adjusted returns, allowing comparison of performance against the risk taken.
Sortino	Sortino ratio: $\mathbb{E}[R]/\text{DD}$.	Focuses on downside risk, offering a refined view of risk-adjusted performance.
MDD	Maximum drawdown	Indicates the largest observed loss from peak to trough, essential for assessing strategy resilience.
Calmar	Calmar ratio: $\mathbb{E}[R]/\text{MDD}$.	Relates annual return to maximum drawdown, providing a perspective on performance during adverse conditions.
% +ve	Percentage of positive trade returns.	Reflects the strategy's consistency in yielding profits, a key aspect for investment decision-making.
Avg+/Avg-	Ratio of average positive to negative trade returns.	Shows the balance between average gains and losses, indicating overall trade effectiveness.

Table 1: Performance Metrics for Algorithm Evaluation

Reward Function: In evaluating performance, we adopted the Differential Sharpe Ratio (DSR) as our reward function. Articulated in the classic Financial RL paper by Moody and Saffell (1998), the DSR can be formulated as follows [146]:

$$D_t = \frac{B_{t-1}\Delta A_t - \frac{1}{2}A_{t-1}\Delta B_t}{(B_{t-1} - A_{t-1}^2)^{\frac{3}{2}}}, \tag{13}$$

where A_t and B_t represent the exponential moving averages of the returns and squared returns, respectively. ΔA_t and ΔB_t are the changes in these averages at time t. This metric, utilizing exponential moving averages, is particularly suited for the dynamic and online nature of RL environments, offering immediate, risk-adjusted performance feedback vital for RL decision-making[9]. The DSR's focus on consistent returns and effective risk management aligns well with our goals of guiding the agent toward risk-adjusted strategies similar to those of a long-only asset manager. Different reward metrics might be preferred for strategies focusing on absolute returns or conservative approaches, emphasizing the need to tailor the reward function to specific investment objectives and regulatory requirements.

Data Selection Rationale: The data selection process for our study was meticulously planned to ensure both the validity and practical applicability of our models. We selected 11 high-volume U.S. equities and indexes from varied sectors including industrials, energy, financials, and consumer goods. This choice was driven by two main factors: Firstly, to minimize market impact and transaction costs. High-volume stocks were selected for their liquidity, which is crucial in real-world trading environments, reducing the potential market impact and costs associated with trading activities. Secondly, our focus was on enabling effective long-term horizon portfolio allocation strategies. These stocks and indexes are well-suited for strategies involving real-world Exchange Traded Funds (ETFs) and individual stocks. By opting for a low cardinality set of assets, as supported by existing literature, we aimed to focus on the sim-to-real dynamics within a controlled yet realistic market setting, avoiding excessive market frictions. Additionally, we started with a moderate initial cash amount and explicitly excluded the use of *leverage* or borrowing, to further reduce market impact and align with real-world trading constraints.

Reinforcement Learning Models and Experimental Results: We have carefully chosen a set of Reinforcement Learning (RL) algorithms that have been demonstrated to surpass traditional investment strategies, such as those based on Modern Portfolio Theory (MPT), as discussed in earlier sections[10]. For clarity, especially for readers well-versed in finance terminology, we reiterate that the term *policy* within this context refers to an *algorithmic investment strategy*:

- **Advantage Actor-Critic (A2C)**: Technically, A2C operates with a dual architecture, combining a policy model (the actor) and a value model (the critic). It simultaneously learns a policy and a value function, balancing immediate rewards with long-term value. This makes A2C efficient in high-dimensional action spaces. Intuitively, think of A2C as having a trader (actor) making decisions, with a risk management team (critic) evaluating and guiding these decisions for optimal outcomes.
- **AlphaOptimizerNet Simplified (AONS)**: Our proprietary AONS algorithm, which is based on more recent developments in the state-of-the-art (SOTA) Artificial Intelligence literature was originally designed for proprietary high-frequency intra-day research. The neural architecture was adapted and simplified for our long-term asset management experiments.
- **Proximal Policy Optimization (PPO)**: PPO is innovative in its approach to updating policies, using a clipped objective function to prevent large, destabilizing updates. It maintains a balance between exploring new strategies and exploiting known ones, crucial in unpredictable financial markets. Intuitively, PPO can be seen as an adaptive trader, constantly refining strategies based on market feedback and risk-return profiles.
- **Soft Actor-Critic (SAC)**: SAC employs a stochastic policy framework and off-policy learning, effectively exploring diverse strategies. This makes it adept at handling market uncertainty and volatility. Intuitively, SAC resembles a trader using probabilistic models to anticipate and adapt to various market conditions, enhancing strategy adaptability.

Table 2 details the performance metrics for each RL agent during this out-of-sample backtesting phase, providing insights into their ability to navigate the complexities of the financial market landscape. Our out-of-sample backtesting

[9]The DSR's denominator, raised to the power of $\frac{3}{2}$, emerges from its first-order expansion, capturing the immediate impact of current returns on the Sharpe Ratio. This formulation is unique to the DSR, specifically designed for online applications [147].

[10]While the detailed technical intricacies of these algorithms are beyond the scope of this paper, the provided intuitive explanations aim to highlight their unique attributes and operational analogies in the context of financial portfolio management. It is vital to note the challenges inherent in replicating financial RL experiments, especially regarding computational resources. Studies from Wang et. al (2021) and others [148]–[151] highlight the importance of advanced computing power, vectorized training, and sophisticated memory buffer systems for scaling experiments and improving computational efficiency. Variations in hardware, software, and statistical experimentation design can lead to different outcomes, even with seemingly similar methodologies, features and model parameters.

yields a nuanced perspective on the applicability of reinforcement learning (RL) models under real-world market constraints. Although RL models are often touted for their potential to surpass classical investment strategies, our findings suggest that their efficacy might be compromised under realistic market conditions. These results preliminarily support our hypothesis that our novel sim-to-real considerations may be crucial for the practical deployment of RL models in financial environments, beyond being merely additive. In particular, our proprietary *AlphaOptimizerNet*

Table 2: Out-of-Sample Backtesting Performance Metrics for RL Agents in Portfolio Optimization

Model	**E[R]**	**std(R)**	**DD**	**Sharpe**	**Sortino**	**MDD**	**Calmar**	**Avg+/Avg-**
A2C	0.1382	0.0152	0.0151	0.7927	1.1738	−0.2276	0.5248	−0.9826
AONS	**0.2309**	0.0161	**0.0157**	**1.1459**	**1.7244**	**−0.2395**	**0.8291**	**−1.0373**
PPO	0.1416	0.0152	0.0152	0.8094	1.1897	−0.2511	0.4876	−0.9841
SAC	0.0115	0.0156	0.0162	0.1810	0.2557	−0.2776	0.0360	−0.9120

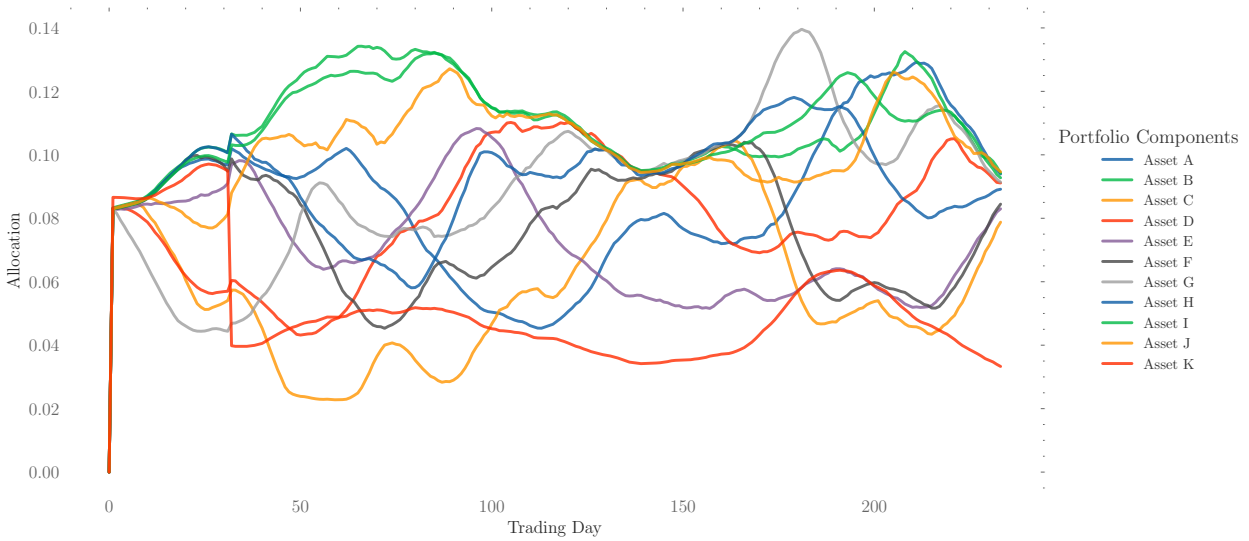

Figure 5: Dynamic portfolio allocation over time by our proprietary *AlphaOptimizerNet*, suggesting a short-term momentum trading strategy. The model's allocation across multiple assets and its responsive adjustments to market conditions hint at its ability to discern meaningful financial patterns and implement complex strategies. Our empirical analysis validates this behavior. For deeper validation, we recommend an extensive examination using modern RL explainability algorithms, as detailed in our novel sim-to-real frameworks section.

Simplified (AONS) has demonstrated significant adaptability within recent market regimes. AONS posted an impressive 23.09% return ($\mathbb{E}[R]$), indicating its strong performance in generating profits. This high return rate is noteworthy, especially when compared with other models like A2C, PPO, and SAC, as it highlights AONS's ability to adapt to market shifts and achieve superior returns. When considering risk-adjusted returns, AONS stands out with a Sharpe ratio of 1.1459 and a Sortino ratio of 1.7244. These metrics suggest that AONS efficiently balances returns with the risks it takes. Moreover, its Calmar ratio of 0.8291, despite a Maximum Drawdown (MDD) of -23.95%, signals a strong recovery potential from losses. It's important to note that MDD, a crucial risk metric, is not annualized. Instead, it measures the largest single drop from peak to trough in the portfolio's value, offering insight into the downside risk over the entire investment period. AONS's negative Avg+/Avg- ratio indicates an aggressive trading strategy, contrasting with the more passive approaches often found in classical Modern Portfolio Theory (MPT). This aggressive approach, focusing on capitalizing on market trends, is a significant departure from traditional strategies and highlights the dynamic nature of RL strategies in finance. Overall, while models like A2C, PPO, and SAC also achieved positive returns, they did not match AONS's high returns and risk-adjusted performance metrics. This underscores the effectiveness of RL in adapting to market changes and outperforming traditional single-period MVO solutions. Figure 5 illustrates AONS's dynamic portfolio allocation strategy, exemplifying the model's proficiency in adjusting to short-term market trends. Such visualizations, coupled with an emphasis on model explainability, are crucial in a field where many financial AI studies tend to neglect the temporal evolution of trading strategies. Our replication attempts of these studies have revealed a trend where models often default to simpler strategies like Uniform Buy and Hold, emphasizing the importance of incorporating realistic market dynamics into financial AI research. Despite

the promising results, they warrant cautious scrutiny, particularly in high-stakes financial domains. The aggressive stance of AONS may elicit concerns within traditional institutional risk management frameworks, reinforcing the need for judicious deployment of such models. The success of AONS, along with the positive outcomes from other RL models, calls for ongoing refinement and empirical validation to ensure their practical applicability in the ever-changing financial market landscape. Our study contributes to the discourse on integrating sophisticated AI techniques into portfolio management, advocating for a balanced approach that fuses technical ingenuity with pragmatic execution.

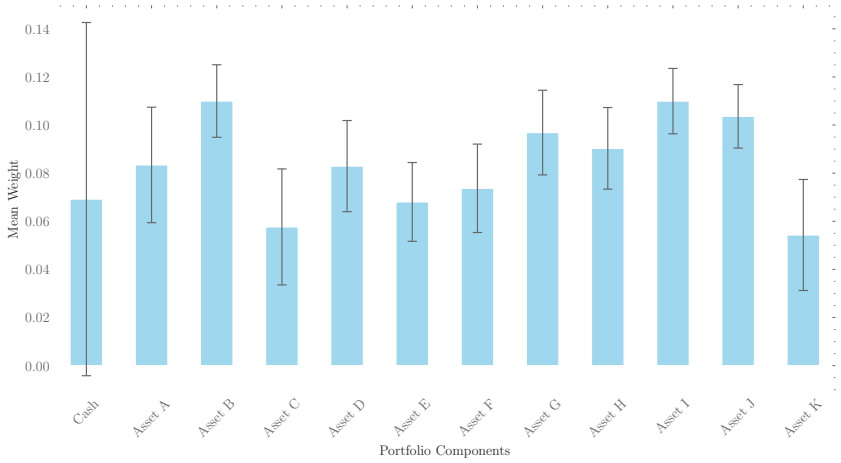

Figure 6: Average allocation per asset by our proprietary *AlphaOptimizerNet* across the testing period, with error bars representing one standard deviation from the mean. This visual representation highlights the model's strategic distribution of investments, with the variance indicating its responsiveness to market conditions. The error bars may extend beyond the mean allocation to suggest the range within which the allocation varied over the period.

5 Conclusion

In conclusion, this paper marks a substantial step towards achieving our initial goals. We have successfully developed the *AlphaOptimizerNet* Reinforcement Learning agent. Furthermore, we have explored the critical aspects of our novel sim-to-real-world transfer frameworks, contributing to the discourse on advanced AI applications in portfolio management. Our preliminary frameworks and extensive computational experiments are designed to serve as a cornerstone for future research and practical implementations in this field. We aspire for our findings and methodologies to elevate financial management standards and inform future explorations. Particularly for financial institutions managing substantial assets, our work offers a methodically designed reference point, potentially enhancing their wealth safeguarding strategies and ability to limit their exposure to risk. Our commitment to advancing this domain remains unwavering, as we continue to innovate in neural network architectures and refine our proposed frameworks. Our ultimate goal is to foster a financial ecosystem that is not only efficient and resilient but also attuned its broader impact on individual prosperity and collective economic growth. We hope that our current and future contributions will significantly *advance investment frontiers*, bridging theoretical research and practical applications.

Disclaimer: The authors make no representation or warranty, express or implied, and disclaim all liability regarding the completeness, accuracy, or reliability of the information contained in this paper. This document is not intended as investment research or advice, nor as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, product, or service.

References

[1] Philip Ndikum. "Machine learning algorithms for financial asset price forecasting". In: *arXiv preprint arXiv:2004.01504* (2020).

[2] Paul Wilmott. "Where quants go wrong: a dozen basic lessons in commonsense for quants and risk managers and the traders who rely on them". In: *Wilmott Journal* 1.1 (2009), pp. 1–22.

[3] David H Bailey and Marcos Lopez de Prado. "Finance is Not Excused: Why Finance Should Not Flout Basic Principles of Statistics". In: *Forthcoming, Significance (Royal Statistical Society)* (2021).

[4] Humphrey K. K. Tung and Michael C. S. Wong. "Financial Risk Forecasting with Non-Stationarity". In: *Financial Risk Forecasting*. Palgrave Macmillan UK, 2011.

[5] Thomas Guhr. "Non-stationarity in Financial Markets: Dynamics of Market States Versus Generic Features". In: *Acta Physica Polonica B* 46 (2015), p. 1625.

[6] Mazin AM Al Janabi. "Optimization algorithms and investment portfolio analytics with machine learning techniques under time-varying liquidity constraints". In: *Journal of Modelling in Management* ahead-of-print (2021).

[7] Marcos López De Prado. "The 10 reasons most machine learning funds fail". In: *The Journal of Portfolio Management* 44.6 (2018), pp. 120–133.

[8] Adrian Millea. "Deep reinforcement learning for trading—A critical survey". In: *Data* 6.11 (2021), p. 119.

[9] Leif Andersen. "Regulation, Capital, and Margining: Quant Angle". In: (2014).

[10] Christos Makridis and Alberto G Rossi. "Rise of the'Quants' in Financial Services: Regulation and Crowding Out of Routine Jobs". In: *Available at SSRN 3218031* (2018).

[11] Sean D Holcomb, William K Porter, Shaun V Ault, et al. "Overview on deepmind and its alphago zero ai". In: *Proceedings of the 2018 international conference on big data and education*. 2018, pp. 67–71.

[12] Johannes Heinrich and David Silver. "Deep reinforcement learning from self-play in imperfect-information games". In: *arXiv preprint arXiv:1603.01121* (2016).

[13] Philip Paquette, Yuchen Lu, Seton Steven Bocco, et al. "No-press diplomacy: Modeling multi-agent gameplay". In: *Advances in Neural Information Processing Systems* 32 (2019).

[14] Noam Brown, Anton Bakhtin, Adam Lerer, et al. "Combining deep reinforcement learning and search for imperfect-information games". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17057–17069.

[15] Jian Yao, Zeyu Zhang, Li Xia, et al. "Solving imperfect information poker games using Monte Carlo search and POMDP models". In: *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE. 2020, pp. 1060–1065.

[16] Anton Bakhtin, David Wu, Adam Lerer, et al. "No-press diplomacy from scratch". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18063–18074.

[17] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, et al. "Deep reinforcement learning for autonomous driving: A survey". In: *IEEE Transactions on Intelligent Transportation Systems* 23.6 (2021), pp. 4909–4926.

[18] Jithin Jagannath, Anu Jagannath, Sean Furman, et al. "Deep learning and reinforcement learning for autonomous unmanned aerial systems: Roadmap for theory to deployment". In: *Deep Learning for Unmanned Systems* (2021), pp. 25–82.

[19] Makram Chahine, Ramin Hasani, Patrick Kao, et al. "Robust flight navigation out of distribution with liquid neural networks". In: *Science Robotics* 8.77 (2023), eadc8892.

[20] Zhiwei Qin, Xiaocheng Tang, Yan Jiao, et al. "Ride-hailing order dispatching at didi via reinforcement learning". In: *INFORMS Journal on Applied Analytics* 50.5 (2020), pp. 272–286.

[21] Terrence J Sejnowski. "Large language models and the reverse turing test". In: *Neural computation* 35.3 (2023), pp. 309–342.

[22] Adaku Uchendu, Zeyu Ma, Thai Le, et al. "TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation". In: *arXiv preprint arXiv:2109.13296* (2021).

[23] Baolin Peng, Chunyuan Li, Pengcheng He, et al. "Instruction tuning with gpt-4". In: *arXiv preprint arXiv:2304.03277* (2023).

[24] Shijie Wu, Ozan Irsoy, Steven Lu, et al. "Bloomberggpt: A large language model for finance". In: *arXiv preprint arXiv:2303.17564* (2023).

[25] Xiao-Yang Liu, Hongyang Yang, Qian Chen, et al. "FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance". In: *arXiv preprint arXiv:2011.09607* (2020).

[26] Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, et al. "FinRL-Meta: Market environments and benchmarks for data-driven financial reinforcement learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1835–1849.

[27] Zechu Li, Xiao-Yang Liu, Jiahao Zheng, et al. "FinRL-Podracer: High performance and scalable deep reinforcement learning for quantitative finance". In: *Proceedings of the Second ACM International Conference on AI in Finance*. 2021, pp. 1–9.

[28] Zitao Song, Xuyang Jin, and Chenliang Li. "Safe-FinRL: A Low Bias and Variance Deep Reinforcement Learning Implementation for High-Freq Stock Trading". In: *arXiv preprint arXiv:2206.05910* (2022).

[29] Antonin Raffin, Ashley Hill, Adam Gleave, et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8. URL: http://jmlr.org/papers/v22/20-1364.html.

[30] Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, et al. "Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms". In: *Journal of Machine Learning Research* 23.274 (2022), pp. 1–18.

[31] Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, et al. *Gymnasium*. Mar. 2023. DOI: 10.5281/zenodo.8127026. URL: https://zenodo.org/record/8127025 (visited on 07/08/2023).

[32] Jiayi Weng, Min Lin, Shengyi Huang, et al. "Envpool: A highly parallel reinforcement learning environment execution engine". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22409–22421.

[33] Ngoc Duy Nguyen, Thanh Thi Nguyen, Nhat Truong Pham, et al. "Towards designing a generic and comprehensive deep reinforcement learning framework". In: *Applied Intelligence* 53.3 (2023), pp. 2967–2988.

[34] Matei Zaharia, Omar Khattab, Lingjiao Chen, et al. *The Shift from Models to Compound AI Systems*. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/. 2024.

[35] Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen. "Is there a replication crisis in finance?" In: *The Journal of Finance* 78.5 (2023), pp. 2465–2518.

[36] Theis Ingerslev Jensen, Bryan T Kelly, and Lasse Heje Pedersen. *Is there a replication crisis in finance?* Tech. rep. National Bureau of Economic Research, 2021.

[37] Kewei Hou, Chen Xue, and Lu Zhang. "Replicating anomalies". In: *The Review of financial studies* 33.5 (2020), pp. 2019–2133.

[38] Florian Echtler and Maximilian Häußler. "Open source, open science, and the replication crisis in HCI". In: *Extended abstracts of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–8.

[39] Matthew Hutson. *Artificial intelligence faces reproducibility crisis*. 2018.

[40] Elizabeth Gibney. "Is AI fuelling a reproducibility crisis in science". In: *Nature* 608 (2022), pp. 250–251.

[41] Yann-Gaël Guéhéneuc and Foutse Khomh. "Empirical software engineering". In: *Handbook of Software Engineering* (2019), pp. 285–320.

[42] Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, et al. "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 335–348.

[43] Christopher Tong. "Statistical inference enables bad science; statistical thinking enables good science". In: *The American Statistician* 73.sup1 (2019), pp. 246–261.

[44] Roger Mead, Steven G Gilmour, and Andrew Mead. *Statistical principles for the design of experiments: applications to real experiments*. Vol. 36. Cambridge University Press, 2012.

[45] Michael Parkinson and Carlos Oscar Sánchez Sorzano. "Why Do We Need a Statistical Experiment Design?" In: *Experimental Design and Reproducibility in Preclinical Animal Studies* (2021), pp. 129–146.

[46] Hartmut Schiefer, Felix Schiefer, Hartmut Schiefer, et al. "Statistical Design of Experiments (DoE)". In: *Statistics for Engineers: An Introduction with Examples from Practice* (2021), pp. 1–20.

[47] Xiaoning Kang and Xinwei Deng. "Design and analysis of computer experiments with quantitative and qualitative inputs: A selective review". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.3 (2020), e1358.

[48] Tyler Munger and Subhas Desa. "A statistical design of experiments approach to machine learning model selection in engineering applications". In: *Journal of Computing and Information Science in Engineering* 21.1 (2021), p. 011008.

[49] Frank J Fabozzi, Francis Gupta, and Harry M Markowitz. "The legacy of modern portfolio theory". In: *The journal of investing* 11.3 (2002), pp. 7–22.

[50] Warren B Powell. *Reinforcement Learning and Stochastic Optimization: A unified framework for sequential decisions*. John Wiley & Sons, 2022.

[51] Zvi Bodie, Alex Kane, and Alan Marcus. *Investments: Global Edition*. McGraw Hill, 2020.

[52] Harry M Markowitz and Harry M Markowitz. *Portfolio selection: efficient diversification of investments*. J. Wiley, 1967.

[53] Franco Modigliani and Merton H Miller. "The cost of capital, corporation finance and the theory of investment". In: *The American economic review* 48.3 (1958), pp. 261–297.

[54] William F Sharpe. "Capital asset prices: A theory of market equilibrium under conditions of risk". In: *The journal of finance* 19.3 (1964), pp. 425–442.

[55] Harry M Markowitz. "Foundations of portfolio theory". In: *The journal of finance* 46.2 (1991), pp. 469–477.

[56] T-J Chang, Nigel Meade, John E Beasley, et al. "Heuristics for cardinality constrained portfolio optimisation". In: *Computers & Operations Research* 27.13 (2000), pp. 1271–1302.

[57] Francesco Cesarone, Andrea Scozzari, and Fabio Tardella. "A new method for mean-variance portfolio optimization with cardinality constraints". In: *Annals of Operations Research* 205 (2013), pp. 213–234.

[58] Dimitris Kouzoupis, Gianluca Frison, Andrea Zanelli, et al. "Recent advances in quadratic programming algorithms for nonlinear model predictive control". In: *Vietnam Journal of Mathematics* 46.4 (2018), pp. 863–882.

[59] Markus Hirschberger, Yue Qi, and Ralph E Steuer. "Large-scale MV efficient frontier computation via a procedure of parametric quadratic programming". In: *European Journal of Operational Research* 204.3 (2010), pp. 581–588.

[60] Stephen Boyd, Enzo Busseti, Steve Diamond, et al. "Multi-period trading via convex optimization". In: *Foundations and Trends® in Optimization* 3.1 (2017), pp. 1–76.

[61] Olga Bourachnikova and Thierry Burger-Helmchen. "Investor's behaviour and the relevance of asymmetric risk measures". In: *Banks & bank systems* 7, Iss. 2 (2012), pp. 87–94.

[62] Borjana Racheva-Iotova and Stoyan Stoyanov. "Post-modern approaches for portfolio optimization". In: *Handbook on information technology in finance*. Springer, 2008, pp. 613–634.

[63] Edmond Lezmi, Thierry Roncalli, and Jiali Xu. "Multi-Period Portfolio Optimization". In: *Available at SSRN* (2022).

[64] Petter N Kolm, Reha Tütüncü, and Frank J Fabozzi. "60 Years of portfolio optimization: Practical challenges and current trends". In: *European Journal of Operational Research* 234.2 (2014), pp. 356–371.

[65] Jean-Philippe Bouchaud. "Economics needs a scientific revolution". In: *Nature* 455.7217 (2008), pp. 1181–1181.

[66] Cristina Geambasu, Robert Sova, Iulia Jianu, et al. "Risk measurement in post-modern portfolio theory: differences from modern portfolio theory". In: *Economic Computation & Economic Cybernetics Studies & Research* 47.1 (2013), pp. 113–132.

[67] Lionel Martellini. "Toward the design of better equity benchmarks: Rehabilitating the tangency portfolio from modern portfolio theory". In: *The Journal of Portfolio Management* 34.4 (2008), pp. 34–41.

[68] Alexander Kempf, Olaf Korn, and Sven Saßning. "Portfolio optimization using forward-looking information". In: *Review of Finance* 19.1 (2015), pp. 467–490.

[69] Daniele D'Alvia. "Uncertainty: The Necessary Unknowable Road to Speculation". In: *The Speculator of Financial Markets: How Financial Innovation and Supervision Made the Modern World*. Cham: Springer International Publishing, 2023, pp. 119–169.

[70] Erdinç Akyıldırım and Halil Mete Soner. "A brief history of mathematics in finance". In: *Borsa Istanbul Review* 14.1 (2014), pp. 57–63.

[71] Samuel Schmidgall, Jascha Achterberg, Thomas Miconi, et al. "Brain-inspired learning in artificial neural networks: a review". In: *arXiv preprint arXiv:2305.11252* (2023).

[72] Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. "A Study of Biologically Plausible Neural Network: The Role and Interactions of Brain-Inspired Mechanisms in Continual Learning". In: *arXiv preprint arXiv:2304.06738* (2023).

[73] Karl Johan Åström. "Optimal control of Markov processes with incomplete state information". In: *Journal of mathematical analysis and applications* 10.1 (1965), pp. 174–205.

[74] Michael L Littman. "A tutorial on partially observable Markov decision processes". In: *Journal of Mathematical Psychology* 53.3 (2009), pp. 119–125.

[75] Richard Bellman. "A Markovian decision process". In: *Journal of mathematics and mechanics* (1957), pp. 679–684.

[76] Huaizhi Wang, Zhenxing Lei, Xian Zhang, et al. "A review of deep learning for renewable energy forecasting". In: *Energy Conversion and Management* 198 (2019), p. 111799.

[77] Alejandro J del Real, Fernando Dorado, and Jaime Durán. "Energy demand forecasting using deep learning: Applications for the French grid". In: *Energies* 13.9 (2020), p. 2242.

[78] Ioannis Boukas, Damien Ernst, Thibaut Théate, et al. "A deep reinforcement learning framework for continuous intraday market bidding". In: *Machine Learning* 110 (2021), pp. 2335–2387.

[79] Yuanhang Zheng, Zeshui Xu, and Anran Xiao. "Deep learning in economics: a systematic and critical review". In: *Artificial Intelligence Review* (2023), pp. 1–43.

[80] Jay Cao, Jacky Chen, John Hull, et al. "Deep learning for exotic option valuation". In: *The Journal of Financial Data Science* (2021).

[81] Anthony L Caterini and Dong Eui Chang. "Generic Representation of Neural Networks". In: *Deep Neural Networks in a Mathematical Framework*. Springer, 2018, pp. 23–34.

[82] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, et al. "Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next". In: *arXiv preprint arXiv:2201.05624* (2022).

[83] James Owen Weatherall. *The physics of wall street: a brief history of predicting the unpredictable*. Houghton Mifflin Harcourt, 2013.

[84] Alex Tuzhilin, Yehuda Koren, Jim Bennett, et al. "Large-scale recommender systems and the netflix prize competition". In: *KDD Proceedings*. 2008, pp. 1–34.

[85] Robert M Bell and Yehuda Koren. "Lessons from the Netflix prize challenge". In: *Acm Sigkdd Explorations Newsletter* 9.2 (2007), pp. 75–79.

[86] Harald Steck, Linas Baltrunas, Ehtsham Elahi, et al. "Deep learning for recommender systems: A Netflix case study". In: *AI Magazine* 42.3 (2021), pp. 7–18.

[87] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, et al. "Deep matrix factorization models for recommender systems." In: *IJCAI*. Vol. 17. Melbourne, Australia. 2017, pp. 3203–3209.

[88] Lucas de Azevedo Takara, André Alves Portela Santos, Viviana Cocco Mariani, et al. "Deep Reinforcement Learning Applied to a Sparse-Reward Trading Environment with Intraday Data". In: *Available at SSRN 4411793* ().

[89] Chen Gao, Yu Zheng, Nian Li, et al. "A survey of graph neural networks for recommender systems: Challenges, methods, and directions". In: *ACM Transactions on Recommender Systems* 1.1 (2023), pp. 1–51.

[90] Zheqing Zhu, Rodrigo de Salvo Braz, Jalaj Bhandari, et al. "Pearl: A Production-ready Reinforcement Learning Agent". In: *arXiv preprint arXiv:2312.03814* (2023).

[91] Andrew Bennett, Nathan Kallus, Lihong Li, et al. "Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1999–2007.

[92] Rachana Mehta and Keyur Rana. "A review on matrix factorization techniques in recommender systems". In: *2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*. IEEE. 2017, pp. 269–274.

[93] Shuai Zhang, Lina Yao, Aixin Sun, et al. "Deep learning based recommender system: A survey and new perspectives". In: *ACM computing surveys (CSUR)* 52.1 (2019), pp. 1–38.

[94] M Mehdi Afsar, Trafford Crump, and Behrouz Far. "Reinforcement learning based recommender systems: A survey". In: *ACM Computing Surveys* 55.7 (2022), pp. 1–38.

[95] Yuanguo Lin, Yong Liu, Fan Lin, et al. "A survey on reinforcement learning for recommender systems". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[96] Saud Almahdi and Steve Y Yang. "An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown". In: *Expert Systems with Applications* 87 (2017), pp. 267–279.

[97] Yoshiharu Sato. "Model-free reinforcement learning for financial portfolios: a brief survey". In: *arXiv preprint arXiv:1904.04973* (2019).

[98] Amine Mohamed Aboussalah and Chi-Guhn Lee. "Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization". In: *Expert Systems with Applications* 140 (2020), p. 112891.

[99] Hui Niu, Siyuan Li, and Jian Li. "MetaTrader: An reinforcement learning approach integrating diverse policies for portfolio optimization". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 1573–1583.

[100] Zihao Zhang, Stefan Zohren, and Roberts Stephen. "Deep reinforcement learning for trading". In: *The Journal of Financial Data Science* (2020).

[101] Leonardo Kanashiro Felizardo, Francisco Caio Lima Paiva, Anna Helena Reali Costa, et al. "Reinforcement Learning Applied to Trading Systems: A Survey". In: *arXiv preprint arXiv:2212.06064* (2022).

[102] Gordon Ritter. "Machine learning for trading". In: *Available at SSRN 3015609* (2017).

[103] Bruno Biais, Larry Glosten, and Chester Spatt. "Market microstructure: A survey of microfoundations, empirical results, and policy implications". In: *Journal of Financial Markets* 8.2 (2005), pp. 217–264.

[104] David Easley, Marcos López de Prado, Maureen O'Hara, et al. "Microstructure in the machine age". In: *The Review of Financial Studies* 34.7 (2021), pp. 3316–3363.

[105] Bastien Baldacci. "Quantitative finance at the microstructure scale: algorithmic trading and regulation". In: (2021).

[106] Tobias Galla and J Doyne Farmer. "Complex dynamics in learning complicated games". In: *Proceedings of the National Academy of Sciences* 110.4 (2013), pp. 1232–1236.

[107] Reidar B Bratvold and Frank Koch. "Game Theory in the Oil and Gas Industry". In: *The Way Ahead* 7.01 (2011), pp. 18–20.

[108] Vivek K Pandey and Chen Y Wu. "Investors May Take Heart: A Game Theoretic View of High Frequency Trading". In: *Journal of Financial Planning* 28.5 (2015), pp. 53–57.

[109] Prajit T Rajendran, Huascar Espinoza, Agnes Delaborde, et al. "Human-in-the-Loop Learning Methods Toward Safe DL-Based Autonomous Systems: A Review". In: *Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops: DECSoS, MAPSOD, DepDevOps, USDAI, and WAISE, York, UK, September 7, 2021, Proceedings 40*. Springer. 2021, pp. 251–264.

[110] Matthew E Taylor. "Reinforcement Learning Requires Human-in-the-Loop Framing and Approaches." In: *HHAI*. 2023, pp. 351–360.

[111] Alvaro HC Correia and Freddy Lecue. "Human-in-the-loop feature selection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 2438–2445.

[112] Jingda Wu, Zhiyu Huang, Chao Huang, et al. "Human-in-the-loop deep reinforcement learning with application to autonomous driving". In: *arXiv preprint arXiv:2104.07246* (2021).

[113] Sebastian Höfer, Kostas Bekris, Ankur Handa, et al. "Perspectives on sim2real transfer for robotics: A summary of the r: Ss 2020 workshop". In: *arXiv preprint arXiv:2012.03806* (2020).

[114] Saminda Wishwajith Abeyruwan, Laura Graesser, David B D'Ambrosio, et al. "i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops". In: *Conference on Robot Learning*. PMLR. 2023, pp. 212–224.

[115] Xiaoyu Chen, Jiachen Hu, Chi Jin, et al. "Understanding domain randomization for sim-to-real transfer". In: *arXiv preprint arXiv:2110.03239* (2021).

[116] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. "Sim-to-real transfer in deep reinforcement learning for robotics: a survey". In: *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE. 2020, pp. 737–744.

[117] Andrei A Rusu, Matej Večerík, Thomas Rothörl, et al. "Sim-to-real robot learning from pixels with progressive nets". In: *Conference on robot learning*. PMLR. 2017, pp. 262–270.

[118] Jiachen Hu, Han Zhong, Chi Jin, et al. "Provable sim-to-real transfer in continuous domain with partial observations". In: *arXiv preprint arXiv:2210.15598* (2022).

[119] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[120] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges". In: *Information fusion* 76 (2021), pp. 243–297.

[121] Owen Lockwood and Mei Si. "A review of uncertainty for deep reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 18. 1. 2022, pp. 155–162.

[122] Yi Zhu, Jing Dong, and Henry Lam. "Uncertainty Quantification and Exploration for Reinforcement Learning". In: *Operations Research* (2023).

[123] Gianluca Bianchin, Yin-Chen Liu, and Fabio Pasqualetti. "Secure navigation of robots in adversarial environments". In: *IEEE Control Systems Letters* 4.1 (2019), pp. 1–6.

[124] Lifeng Zhou and Pratap Tokekar. "Multi-robot coordination and planning in uncertain and adversarial environments". In: *Current Robotics Reports* 2 (2021), pp. 147–157.

[125] Josh Tobin, Rachel Fong, Alex Ray, et al. "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2017, pp. 23–30.

[126] Georgios D Kontes, Daniel D Scherer, Tim Nisslbeck, et al. "High-speed collision avoidance using deep reinforcement learning and domain randomization for autonomous vehicles". In: *2020 IEEE 23rd international conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2020, pp. 1–8.

[127] Quan Vuong, Sharad Vikram, Hao Su, et al. "How to pick the domain randomization parameters for sim-to-real transfer of reinforcement learning policies?" In: *arXiv preprint arXiv:1903.11774* (2019).

[128] Kimin Lee, Kibok Lee, Jinwoo Shin, et al. "Network randomization: A simple technique for generalization in deep reinforcement learning". In: *arXiv preprint arXiv:1910.05396* (2019).

[129] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, et al. "Offline reinforcement learning: Fundamental barriers for value function approximation". In: *arXiv preprint arXiv:2111.10919* (2021).

[130] Tengyang Xie, Nan Jiang, Huan Wang, et al. "Policy finetuning: Bridging sample-efficient offline and online reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 27395–27407.

[131] Tengyang Xie, Dylan J Foster, Yu Bai, et al. "The role of coverage in online reinforcement learning". In: *arXiv preprint arXiv:2210.04157* (2022).

[132] Lindsay Wells and Tomasz Bednarz. "Explainable ai and reinforcement learning—a systematic review of current approaches and trends". In: *Frontiers in artificial intelligence* 4 (2021), p. 550030.

[133] Prashan Madumal, Tim Miller, Liz Sonenberg, et al. "Explainable reinforcement learning through a causal lens". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 03. 2020, pp. 2493–2500.

[134] Jasmina Gajcin and Ivana Dusparic. "Redefining Counterfactual Explanations for Reinforcement Learning: Overview, Challenges and Opportunities". In: *ACM Computing Surveys* (2024).

[135] Petar Kormushev, Sylvain Calinon, and Darwin G Caldwell. "Reinforcement learning in robotics: Applications and real-world challenges". In: *Robotics* 2.3 (2013), pp. 122–148.

[136] Julian Ibarz, Jie Tan, Chelsea Finn, et al. "How to train your robot with deep reinforcement learning: lessons we have learned". In: *The International Journal of Robotics Research* 40.4-5 (2021), pp. 698–721.

[137] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, et al. "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis". In: *Machine Learning* 110.9 (2021), pp. 2419–2468.

[138] Henry Zhu, Justin Yu, Abhishek Gupta, et al. "The ingredients of real-world robotic reinforcement learning". In: *arXiv preprint arXiv:2004.12570* (2020).

[139] Avi Singh, Larry Yang, Kristian Hartikainen, et al. "End-to-end robotic reinforcement learning without reward engineering". In: *arXiv preprint arXiv:1904.07854* (2019).

[140] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, et al. "Scaling data-driven robotics with reward sketching and batch reinforcement learning". In: *arXiv preprint arXiv:1909.12200* (2019).

[141] Abhishek Gupta, Aldo Pacchiano, Yuexiang Zhai, et al. "Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15281–15295.

[142] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, et al. "Inverse reward design". In: *Advances in neural information processing systems* 30 (2017).

[143] Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, et al. "Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning". In: *arXiv preprint arXiv:2004.10190* (2020).

[144] Lukas Brunke, Melissa Greeff, Adam W Hall, et al. "Safe learning in robotics: From learning-based control to safe reinforcement learning". In: *Annual Review of Control, Robotics, and Autonomous Systems* 5 (2022), pp. 411–444.

[145] Bryan Lim, Stefan Zohren, and Stephen Roberts. "Enhancing time-series momentum strategies using deep neural networks". In: *The Journal of Financial Data Science* (2019).

[146] John Moody, Lizhong Wu, Yuansong Liao, et al. "Performance functions and reinforcement learning for trading systems and portfolios". In: *Journal of forecasting* 17.5-6 (1998), pp. 441–470.

[147] Zhiyu Zhang, David Bombara, and Heng Yang. "Discounted Adaptive Online Prediction". In: *arXiv preprint arXiv:2402.02720* (2024).

[148] Zhicheng Wang, Biwei Huang, Shikui Tu, et al. "DeepTrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions Embedding". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 1. 2021, pp. 643–650.

[149] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. "A deep reinforcement learning framework for the financial portfolio management problem". In: *arXiv preprint arXiv:1706.10059* (2017).

[150] Pengqian Yu, Joon Sern Lee, Ilya Kulyatin, et al. "Model-based deep reinforcement learning for dynamic portfolio optimization". In: *arXiv preprint arXiv:1901.08740* (2019).

[151] Srijan Sood, Kassiani Papasotiriou, Marius Vaiciulis, et al. "Deep Reinforcement Learning for Optimal Portfolio Allocation: A Comparative Study with Mean-Variance Optimization". In: *FinPlan 2023* (2023), p. 21.