

Multi-Agent Reinforcement Learning Correctable Strategy: A Framework with Correctable Strategies for Portfolio Management [†]

Kuang-Da Wang ¹, Pei-Xuan Li ² , Hsun-Ping Hsieh ^{2,3,*}  and Wen-Chih Peng ¹

¹ Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan; gdwang.cs10@nycu.edu.tw (K.-D.W.); wcpeng@cs.nycu.edu.tw (W.-C.P.)

² Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan; n28121107@gs.ncku.edu.tw

³ Academy of Innovative Semiconductor and Sustainable Manufacturing, National Cheng Kung University, Tainan 70101, Taiwan

* Correspondence: hphsieh@mail.ncku.edu.tw

[†] Presented at 8th International Conference on Knowledge Innovation and Invention 2025 (ICKII 2025), Fukuoka, Japan, 22–24 August 2025.

Abstract

Portfolio management (PM) is a broad investment strategy aimed at risk mitigation through diversified financial product investments. Acknowledging the significance of dynamic adjustments after establishing a portfolio to enhance stability and returns, we propose employing reinforcement learning (RL) to address dynamic decision-making challenges. However, traditional RL methods often struggle to adapt to significant market volatility, primarily by focusing on adjusting existing asset weights. Different from traditional RL methods, the multi-agent reinforcement learning correctable strategy (MAC) developed in this study detects and replaces potentially harmful assets with familiar alternatives, ensuring a resilient response to market crises. Utilizing the multi-agent reinforcement learning model, MAC empowers individual agents to maximize portfolio returns and minimize risk separately. During training, MAC strategically replaces assets to simulate market changes, allowing agents to learn risk-identification through uncertainty estimation. During testing, MAC detects potentially harmful assets and replaces them with more reliable alternatives, enhancing portfolio stability. Experiments conducted on a real-world US Exchange-Traded Fund (ETF) market dataset demonstrate MAC's superiority over standard RL-based PM methods and other baselines, underscoring its practical efficacy for real-world applications.

Keywords: reinforcement learning; investment strategy; uncertainty estimation; portfolio management



Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 29 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Portfolio management (PM) optimizes returns and risks by strategically distributing assets [1]. With limited funds, investors must carefully construct portfolios based on historical performance. While an initial robust portfolio is crucial, asset performance fluctuations over the investment horizon elevate risk, making dynamic portfolio adjustments essential. This work addresses limited investment options, exploring how to improve risk control by dynamically adjusting asset weights and replacing underperforming assets.

Reinforcement learning (RL) has shown its effectiveness in portfolio management due to its adaptability and reward-driven decision-making [2,3]. However, existing studies often focus solely on maximizing returns [2,4], neglecting constraints such as limited assets and capital. Moreover, they lack mechanisms to replace harmful assets under volatile market conditions.

In this paper, we propose a multi-agent reinforcement learning correctable strategy (MAC) for portfolio management to address the above issues. MAC employs multi-agents to optimize various performance metrics, including annual return and the Sharpe ratio, leveraging the strengths of multi-agent systems in handling multifaceted tasks.

During training, we propose the portfolio disruptor (PD) to strategically replace assets in the portfolio, simulating future market changes and maximizing each agent's uncertainty regarding non-original assets. During testing, we introduce the portfolio corrector (PC) to enable agents to dynamically identify potentially harmful or unfamiliar assets and replace them with those in which agents have higher confidence. This enhances portfolio stability by avoiding risky decisions in uncertain situations.

In this study, we introduce MAC for portfolio management by replacing assets and adjusting weights. We incorporated uncertainty estimation into MAC, facilitating the dynamic replacement of harmful assets with better alternatives, thereby improving portfolio stability during market fluctuations. Experiments with real datasets indicate that MAC outperforms existing traditional and RL-based PM methods.

2. Related Work

2.1. RL in Portfolio Investment

Recent RL research has shown heightened interest in quantitative trading and portfolio management [5–8]. This trend is fueled by RL's capacity to train agents to yield profitable trading actions directly, enhancing their generalization across diverse market conditions [9] compared to traditional methods [10,11]. Researchers have explored multi-agent reinforcement learning (MARL) approaches, such as the cooperative system [12], where each agent constructs its portfolio for diversification.

2.2. Uncertainty Estimation in Deep Learning

In RL, two common methods for estimating uncertainty are training independent networks with an ensemble strategy to derive a distribution over predictions [13] and using variance networks to predict mean and variance directly from the input [14]. In this study, we used variance networks, which enhance training speed by eliminating the need for sampling, to estimate uncertainty in portfolio management.

3. Background Knowledge

3.1. Financial Concepts

Annual return (AR) is a common measure of investment performance that considers the net worth and the period length. Annual standard deviation (AStd) can indicate the degree of volatility of assets. It measures the magnitude of return fluctuations caused by the portfolio's factors. Sharpe ratio (SR) [15] is a vital metric that signifies the risk-adjusted return of a portfolio. It is derived by dividing the excess return by the standard deviation of the investment's excess returns. These metrics offer a holistic assessment of a portfolio's performance, encompassing both returns and volatility.

3.2. Objective of Portfolio Management

Given a portfolio of n assets over a time horizon T , the portfolio is defined by a sequence of weight vectors $w_t = (w_{1,t}, w_{2,t}, \dots, w_{n,t})$ where $w_{i,t}$ denotes the allocation of

asset i at time t . The portfolio state at time t is $s_t = (v_{1,t}, v_{2,t}, \dots, v_{n,t})$ where $v_{i,t}$ is the current value of asset i . The objective of portfolio management is to dynamically adjust w_t and replace underperforming assets to maximize AR and SR.

4. Methodology

We introduce a novel portfolio management strategy, MAC (Multi-Agent Reinforcement Learning Correctable Strategy). As shown in Figure 1, MAC consists of three components: a reward function incorporating uncertainty, PD, and PC. These elements leverage two deep deterministic policy gradient (DDPG) [16] agents, chosen for their suitability for this task due to their ability to handle continuous action spaces.

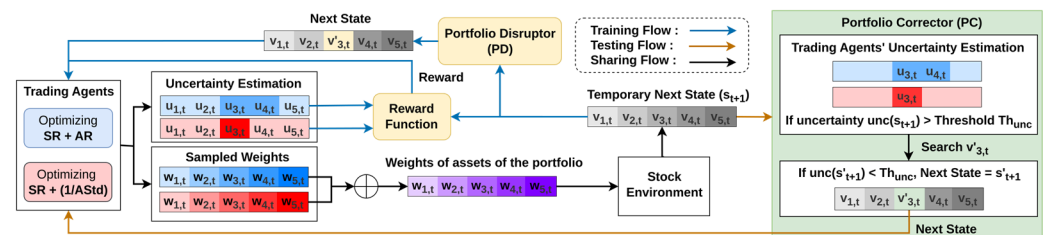


Figure 1. Architecture of the MAC.

4.1. Reward Function

Investors evaluate two pivotal indicators (AR and SR) in portfolio management. To balance return and risk, we train two agents with separate objectives, mainly AR and AStd, instead of adopting a joint multi-objective reinforcement learning (MORL) model, which is often challenging and time-consuming [17]. Both agents' reward functions also incorporate SR to prevent extreme behaviors. In addition, the financial markets are sensitive to a variety of influences that often lead to substantial fluctuations in asset prices. To ensure the agent's decision-making is confident, we introduce the uncertainty estimation into the reward function by a probabilistic neural network (PNN) [18].

PNN assumes a conditional Gaussian for predictive distribution, estimating the mean and standard deviation, which together form a probability distribution function to capture uncertainty effectively. Activating the actor network in the agent with a PNN allows the generation of distinct Gaussian distributions. Each distribution, corresponding to an asset, evaluates the uncertainty in the asset's weight. Interpreting the standard deviations within these Gaussian distributions as asset uncertainties, distributions with lower uncertainty are more concentrated. This concentration signifies the agent's heightened confidence in adjusting asset proportions.

Given that PD within MAC sells assets within the portfolio to purchase other assets, mimicking market dynamics, agents utilize uncertainty estimation to identify non-portfolio assets during training. During training, the objective is to minimize assets' uncertainty within the original portfolio and maximize the uncertainty of assets outside the original portfolio. Consequently, we incorporate uncertainty into the reward function as shown in Equations (1) and (2).

$$r_A(s_t, w_t, s_{t+1}) = SR(s_{t+1}) + AR(s_{t+1}) - unc(s_{t+1}), \quad (1)$$

$$r_B(s_t, w_t, s_{t+1}) = SR(s_{t+1}) + \left(\frac{1}{AStd(s_{t+1})} \right) - unc(s_{t+1}), \quad (2)$$

where $unc(s_{t+1})$ is the uncertainty of the agent for the state s_{t+1} . r_A and r_B represent the reward functions of the two agents, with $SR(s_{t+1})$, $AR(s_{t+1})$ and $AStd(s_{t+1})$ denoting Sharpe ratio, annual return, and annual standard deviation at time $t+1$, respectively. This

design enables agents to learn uncertainty-aware strategies; elevated uncertainty reflects unfamiliar assets and potential risk, which can be addressed by PC during testing.

4.2. Portfolio Disruptor in the Training Stage

We explored the relationship between the uncertainty in the MAC's reward function and the uncertainty of assets. Assets in the original portfolio tend to have lower uncertainty, while others show higher uncertainty. To help agents learn to assess asset uncertainty, PD randomly replaces original assets with unfamiliar ones during training. As illustrated in Figure 2, for each episode of length T , PD determines the disruption count D , indicating how many times it will disrupt the investment portfolio within one episode. At each $\frac{T}{D}$ time step, PD randomly selects a time point $Dt \in (t, t + \frac{T}{D})$ and replaces portfolio assets with random alternatives to simulate market changes, helping agents learn to distinguish between familiar and unfamiliar assets.

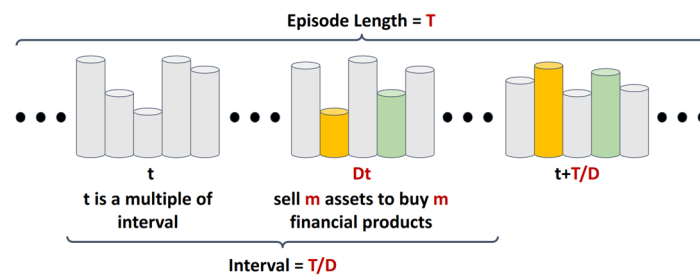


Figure 2. Illustration of PD simulating market changes by disrupting portfolios.

4.3. Portfolio Corrector in Testing

After training, we derive an investment strategy, but testing may still encounter unfamiliar market fluctuations, leading to decisions without confidence. Instead of enabling the agent to handle unknown situations, we developed a correctable strategy utilizing PC to replace potentially harmful assets, avoiding making risky decisions in unknown situations, enhancing stability, and preserving the effectiveness of learned strategies. As shown in Figure 3, after adjusting portfolio weights, the trading agents obtain a temporary next state s_{t+1} , then the PC estimates its uncertainty $unc(s_{t+1})$. If both agents' uncertainties exceed the threshold Th_{unc} , PC greedily searches for an alternate asset satisfying $unc(s'_{t+1}) < Th_{unc}$. If found, it updates the portfolio accordingly, resulting in a new state s'_{t+1} as the next state; otherwise, it retains the original allocation and returns s_{t+1} .

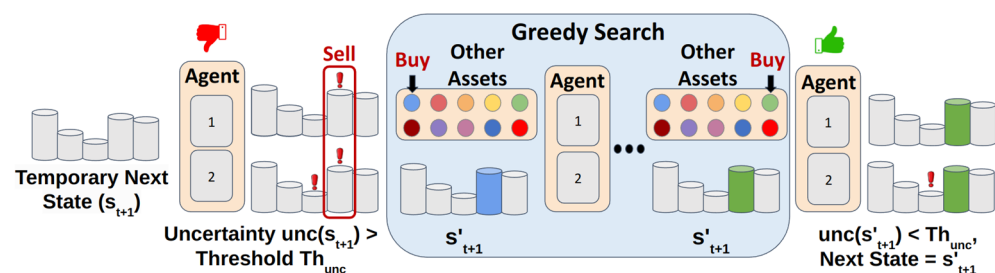


Figure 3. Illustration of PC mitigating portfolio uncertainty during testing.

5. Experiment Results

5.1. Experiment

5.1.1. Dataset

We utilized Yahoo's finance API to acquire U.S. Exchange-Traded Fund (ETF) market data from 1 January 2009 to 31 December 2022. Considering financial constraints, we

limited the portfolio to five assets, aligning with practices of well-known investors, such as the Golden Butterfly, David Swensen, and All Weather portfolios, which typically include four to six components. For a comprehensive evaluation, we randomly selected five ETFs from the 25 largest based on market capitalization, resulting in a total of eight distinct portfolios. Figure 4 presents selected portfolios and their performance during training, showing the variety of scenarios covered by our experiments.

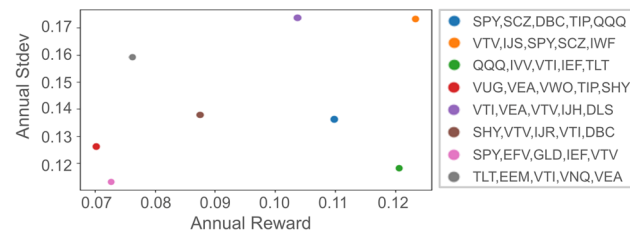


Figure 4. Eight selected portfolios' performances of Buy and Hold in training.

5.1.2. Evaluation Metrics and Baselines

We selected AR and SR as metrics. AR evaluates return, while SR comprehensively covers the risk-return relationship, making them widely used metrics in investment portfolio evaluation. We compared MAC with the following methods: traditional strategies Buy and Hold (B&H) [11] and Mean-Reversion (MR) [10]; traditional RL methods A2C [19] and DDPG [16]; an action-constrained RL (ACRL) method DDPG w/Proj; the state-of-the-art ACRL method NFWPO [20]; the state-of-the-art MORL method GPI-PD [21]; a Transformer-based DRL method RAT [8]; and DRL-based policy optimization (DRPO) [2], which is a state-of-the-art DRL-based method in stock trading.

5.1.3. Implementation Details

Traditional RL and ACRL methods utilize the cumulative asset value as a reward function, akin to DRPO. To explore the impact of different reward functions, we also employ SR as the reward function for DDPG, denoted as DDPG (SR). As for MORL methods, we use both AR and SR objectives to balance reward and risk. In addition, the action space of the ACRL methods is defined as the total permissible operations, with the constraint allowing only five ETF weights to exceed zero. Similar to MAC, the ACRL method can invest in products not originally in the portfolio. For fairness, ten ETFs are randomly selected from the thirty largest by market capitalization and combined with the existing five, resulting in a 15-dimensional action space. For MAC, all the operations are also limited to these 15 ETFs. This process is repeated five times, and average values are presented for a more representative result.

5.1.4. Training and Evaluation Details

The training and validation period spans from 1 January 2009 to 31 December 2019. The testing covers 1 January 2020 to 31 December 2022. Portfolio adjustments incur a 0.1% transaction fee to simulate real trading conditions [5]. Agent training involves 2048 steps per update, with an entropy coefficient of 0.005, a learning rate of 0.0001, and a batch size of 128. We utilize the standard deviation of the output distribution of the PNN as uncertainty, with an uncertainty threshold Th_{unc} set at 1.5. In addition, we randomly selected seeds to repeat the experiment five times and computed their average value to obtain more reliable results.

5.2. Quantitative Results

Table 1 shows that MAC outperforms the baselines across all metrics, emphasizing its applicability in portfolio management. The observations are summarized as follows.

1. Superior Performance of MAC

Traditional strategies (B&H and MR) follow simple rules and fail in crisis. Traditional RL methods (A2C and DDPG) prioritize specific objectives, limiting flexibility. The MORL approach (GPI-PD) aims to address multi-objective problems but encounters challenges like convergence issues. Despite being state-of-the-art, RAT and DRPO focus solely on return or cumulative portfolio value and fail to surpass MAC's AR performance. In contrast, MAC handles conflicting objectives and mitigates risk through the correctable strategy, enabling robust performance across conditions.

2. Correctable strategy advantages

ACRL methods (NFWPO and DDPG w/Proj) maintain stable performance in the experimental portfolio by identifying and allocating higher weights to stable ETFs. However, their frequent asset replacements incur higher transaction costs, leading to reduced returns. In contrast, MAC uses a correctable strategy, enabling agents to assess uncertainty and replace assets only when uncertainty surpasses a predefined threshold, thereby identifying confident assets.

Table 1. Results of all methods across various portfolios are presented as percentages (%). The best result within each row is highlighted in bold, while the second-best result is underlined.

Portfolio	Metric	B&H	MR	A2C	DDPG	DDPG (SR)	GPI-PD	NFWPO	DDPG w/Proj	RAT	DRPO	MAC w/o PC	MAC w/o PC, PD	MAC
SPY, SCZ, DBC, TIP, QQQ	AR	6.3073	5.2591	5.3696	7.8161	7.9308	<u>8.1908</u>	2.3199	2.5451	5.1439	4.3633	−0.8325	5.9714	23.5783
	SR	40.7696	46.9837	30.5771	50.4938	40.4827	<u>54.3957</u>	22.3997	23.566	2.1237	1.6782	5.8484	41.8294	60.2654
VTV, IJS, SPY, SCZ, IWF	AR	4.8479	<u>5.5642</u>	3.9637	3.3871	3.3871	3.2624	3.2774	3.3404	2.675	4.9778	−0.1267	4.9675	45.9088
	SR	31.4051	<u>51.155</u>	15.8451	25.91	13.6535	25.4015	26.5885	26.6724	1.5541	1.8218	11.4429	32.0431	84.557
QQQ, IVV, VTI, IEF, TLT	AR	1.369	1.408	−0.0381	0.2092	0.2092	1.5762	<u>2.9542</u>	2.7248	−0.8771	2.8686	−1.3335	1.5832	4.281
	SR	17.5737	17.2493	−0.234	8.3113	1.546	18.024	<u>25.4625</u>	24.2073	0.2625	1.3118	−1.2405	17.8185	29.1168
VUG, VEA, VWO, TIP, SHY	AR	−1.5585	0.1082	−0.9928	0.5642	0.1848	0.9462	2.6523	<u>2.8634</u>	−0.9454	0.487	1.7259	0.2012	8.0414
	SR	2.2846	<u>28.4078</u>	−6.9272	11.5013	1.3375	13.7621	23.1816	24.3469	−0.0467	0.487	18.5849	8.4538	42.1418
VTI, VEA, VTV, IJH, DLS	AR	2.1391	0.4793	1.5385	0.5721	2.0333	−0.6003	<u>3.6573</u>	3.3548	1.2421	3.6068	2.1474	1.8017	6.2167
	SR	22.0369	14.3403	6.3496	14.3765	8.3425	9.2873	<u>28.2092</u>	26.9288	1.08	1.4624	20.6637	19.5755	37.5446
SPY, EFV, GLD, IEF, VTV	AR	2.5132	0.7561	2.7956	1.3719	1.3719	2.4879	3.5201	3.2095	1.7212	<u>3.6277</u>	−1.2832	3.0242	35.6018
	SR	23.3345	15.304	22.0758	16.5859	8.9795	23.7881	<u>29.662</u>	27.607	1.2205	2.2496	−1.6849	27.2716	71.1542
TLT, EEM, VTI, VNQ, VEA	AR	−3.9244	2.1755	−4.8659	−2.6607	−1.5176	−1.5299	2.1769	2.3197	<u>4.1439</u>	−1.9348	0.7686	−2.5725	4.5716
	SR	−3.8105	<u>26.3245</u>	−28.414	−1.6601	−6.9465	−4.0373	20.8494	21.6167	2.1237	−1.1909	13.5561	−4.2144	31.6202
SHY, VTV, IJR, VTI, DBC	AR	7.8434	2.5765	<u>8.4215</u>	7.1049	5.1871	6.7647	3.8063	3.4638	8.4134	4.2783	2.3736	7.4011	8.8566
	SR	18.9595	31.7973	41.1978	42.1009	29.0448	45.312	28.9443	27.4883	6.2526	1.9444	22.8331	<u>48.3311</u>	50.2616

5.3. Ablation Study

To study the effect of the components in MAC, we also conduct ablation experiments. We compare MAC with the following methods: (1) MAC w/o PC—this variant cannot replace the components in the portfolio; (2) MAC w/o PC, PD—this variant is equal to a MARL featuring two DDPG agents. The reward functions are the same as for MAC, but without considering uncertainty. The results show that MAC exhibits unstable performance without PC and moderate performance when B&H's AR is negative, but it declines significantly when B&H's AR is positive. This arises from the interaction between PC and PD. In stable portfolio situations, PD continuously changes the portfolio during training, resulting in counterproductive outcomes during testing. Conversely, in unstable portfolio situations, the agent operates more confidently. Additionally, there is a notable performance gap between MAC w/o PC, PD, and the complete MAC model, highlighting the effectiveness of PD and PC.

6. Conclusions

Using MAC, we tailored portfolio management under finite asset constraints. The meticulously designed reward function enables agents to optimize various metrics and

learn uncertainty estimation for assets. MAC strategically replaces assets during training to mimic market changes and maximize uncertainty regarding non-original assets. This enables the replacement of potentially harmful assets with more confident alternatives during testing, consistently applying learned strategies from training. Experimental results on real-world datasets underscore MAC's effectiveness, confirming its practicality as a promising strategy for adaptive portfolio management.

Author Contributions: Conceptualization, K.-D.W. and P.-X.L.; methodology, K.-D.W. and P.-X.L.; validation, K.-D.W. and P.-X.L.; formal analysis, K.-D.W. and P.-X.L.; investigation, K.-D.W. and P.-X.L.; resources, H.-P.H. and W.-C.P.; data curation, K.-D.W. and P.-X.L.; writing—original draft preparation, K.-D.W. and P.-X.L.; writing—review and editing, K.-D.W. and P.-X.L.; visualization, K.-D.W. and P.-X.L.; supervision, H.-P.H. and W.-C.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Science and Technology Council (NSTC) of Taiwan under grants 112-2221-E-006-150-MY3 and 114-2628-E-006-005-MY3.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data was retrieved via the Yahoo Finance Python API, and the specific date ranges are described in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Markowitz, H.M. Foundations of portfolio theory. *J. Financ.* **1991**, *46*, 469–477. [\[CrossRef\]](#)
2. Han, L.; Ding, N.; Wang, G.; Cheng, D.; Liang, Y. Efficient Continuous Space Policy Optimization for High-frequency Trading. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023.
3. Liu, X.Y.; Yang, H.; Chen, Q.; Zhang, R.; Yang, L.; Xiao, B.; Wang, C.D. FinRL: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance. *arXiv* **2020**, arXiv:2011.09607. [\[CrossRef\]](#)
4. Liu, X.Y.; Xia, Z.; Rui, J.; Gao, J.; Yang, H.; Zhu, M.; Wang, C.D.; Wang, Z.; Guo, J. FinRL-Meta: Market Environments and Benchmarks for Data-Driven Financial Reinforcement Learning. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022.
5. Sun, S.; Xue, W.; Wang, R.; He, X.; Zhu, J.; Li, J.; An, B. DeepScalper: A Risk-Aware Reinforcement Learning Framework to Capture Fleeting Intraday Trading Opportunities. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, 17–21 October 2022.
6. Wang, J.; Zhang, Y.; Tang, K.; Wu, J.; Xiong, Z. AlphaStock: A Buying-Winners-and-Selling-Losers Investment Strategy using Interpretable Deep Reinforcement Attention Networks. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, Anchorage, AK, USA, 4–8 August 2019.
7. Wang, Z.; Huang, B.; Tu, S.; Zhang, K.; Xu, L. DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021.
8. Xu, K.; Zhang, Y.; Ye, D.; Zhao, P.; Tan, M. Relation-Aware Transformer for Portfolio Policy Learning. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021.
9. Sun, S.; Wang, R.; An, B. Reinforcement learning for quantitative trading. *ACM Trans. Intell. Syst. Technol.* **2023**, *14*, 1–29. [\[CrossRef\]](#)
10. Summers, L.H.; Poterba, J.M. Mean reversion in stock prices: Evidence and implications. *J. Financ. Econ.* **1988**, *22*, 27–59.
11. Shiryayev, A.; Xu, Z.; Zhou, X.Y. Thou shalt buy and hold. *Quant. Financ.* **2008**, *8*, 765–776. [\[CrossRef\]](#)
12. Lee, J.; Kim, R.; Yi, S.W.; Kang, J. MAPS: Multi-Agent reinforcement learning-based Portfolio management System. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021.
13. Lütjens, B.; Everett, M.; How, J.P. Safe RL With Model Uncertainty Estimates. In Proceedings of the International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, 20–24 May 2019.
14. Chung, Y.; Char, I.; Guo, H.; Schneider, J.; Neiswanger, W. Uncertainty Toolbox: An Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification. *arXiv* **2021**, arXiv:2109.10254. [\[CrossRef\]](#)

15. Sharpe, W.F. The Sharpe ratio. *J. Portf. Manag.* **1994**, *21*, 49–58. [[CrossRef](#)]
16. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
17. Liu, C.; Xu, X.; Hu, D. Multiobjective RL: A Comprehensive Overview. *IEEE Trans. Syst. Man Cybern. Syst.* **2015**, *45*, 385–398.
18. Specht, D.F. Probabilistic neural networks. *Neural Netw.* **1990**, *3*, 109–118. [[CrossRef](#)]
19. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.P.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for Deep RL. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
20. Lin, J.L.; Hung, W.; Yang, S.H.; Hsieh, P.C.; Liu, X. Escaping from zero gradient: Revisiting action-constrained reinforcement learning via Frank-Wolfe policy optimization. In Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence, Virtual Event, 27–30 July 2021.
21. Alegre, L.N.; Bazzan, A.L.; Roijers, D.M.; Nowé, A.; da Silva, B.C. Sample-Efficient Multi-Objective Learning via Generalized Policy Improvement Prioritization. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, UK, 29 May–2 June 2023.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.