Universidad Pontificia Comillas
ICAI School of Engineering
Master in Industrial Engineering

Analysis and Application of Reinforcement Learning to Portfolio Optimization

Master's Thesis

Author: Jorge González Calvo
Director: Lucía Güitta López
Co-director: Eduardo César Garrido Merchán

Madrid, 2026

Declaration of originality

I declare under my responsibility that the Project presented with the title **Analysis and Application of Reinforcement Learning to Portfolio Optimization** at the ICAI School of Engineering of the Comillas Pontifical University in the academic year 2025/2026 is of my authorship and has not been presented previously for other purposes. The Project is not plagiarised from any other, either totally or partially, and the information that has been taken from other documents is duly referenced.

Use of Artificial Intelligence[1]

I declare under my responsibility that (indicate the correct option):

☐ I have not used Artificial Intelligence in the preparation of this document.

☑ I have used Artificial Intelligence in the preparation of this document and/or Annex B under the conditions allowed by Comillas Pontifical University, i.e. applying Level 2 of the Perkins et al. (2024) Assessment Scale: "AI can be used for pre-task activities such as brainstorming, description and initial research. This level focuses on the use of AI for planning, synthesising and generating ideas, but assessments should emphasise the ability to develop and refine these ideas independently". Specifically, Artificial Intelligence has been used to:

Artificial Intelligence was used as a support tool during the preparation of the document, strictly adhering to pre-task activities and planning. Specifically, it was used to: brainstorm initial ideas and outline the structural flow of several sections; assist with LaTeX/TikZ syntax and troubleshooting for standard schematic diagrams and document formatting (table layouts, figure margins, and overall structure); suggest potential bibliographic references for the state-of-the-art chapter, which were subsequently located, read, and verified by the author; and help organize and structure the initial concepts for the Sustainable Development Goals section based on the author's raw notes. The AI was not used to draft or edit the final prose of the document. All conceptual development, technical analysis, implementation of the reinforcement learning models, and the final writing were executed entirely independently by the author, who assumes full responsibility for the final document.

[1] This declaration refers to the use of generative Artificial Intelligence to carry out the Project documents (Annex B and Memory). It does not apply to Projects where, by their nature, artificial intelligence must be used as part of them (application of machine learning techniques, neural networks, data analysis...).


Signature (student):
Date: 03/07/2026

Authorisation for Project delivery

Thesis supervisor	Thesis deputy supervisor (if any)
	
Signature:	Signature:
Date: 03/07/2026	Date: 07/07/2026

Acknowledgements

I would like to express my sincere gratitude to my director, Lucía Güitta López, and co-director, Eduardo César Garrido Merchán, for their guidance, availability, and constructive feedback throughout this project. Their expertise in machine learning and financial engineering has been invaluable in shaping both the technical direction and the academic rigour of this work.

I would also like to thank the ICAI School of Engineering for providing the academic environment and resources that made this research possible.

Finally, I am deeply grateful to my family and friends for their constant encouragement and support.

Universidad Pontificia Comillas
ICAI School of Engineering
Master in Industrial Engineering

ANÁLISIS Y APLICACIÓN DEL APRENDIZAJE POR REFUERZO A LA OPTIMIZACIÓN DE CARTERAS

Autor: Jorge González Calvo
Director: Lucía Güitta López
Co-Director: Eduardo César Garrido Merchán

Resumen

Este trabajo investiga la aplicación del Aprendizaje por Refuerzo (RL) a la gestión de carteras de renta variable, centrándose en el algoritmo Proximal Policy Optimization (PPO) dentro del marco FinRL. Se evalúa el comportamiento del agente en dos regímenes de mercado opuestos, un mercado bajista (2022) y uno alcista (2023), sobre un universo de diez activos internacionales entrenado con datos históricos de 2015 a 2021. El estudio se organiza en cinco fases experimentales: una referencia pasiva de compra y mantenimiento, un agente PPO base, un entorno consciente del riesgo con reglas estáticas de stop-loss, la optimización bayesiana de hiperparámetros mediante Optuna y, como contribución original, una función de recompensa basada en el ratio de Sharpe de momentos fraccionarios, motivada por la teoría reciente de López de Prado et al. (2026) sobre la invalidez de la inferencia clásica bajo dinámicas GARCH y colas pesadas. Los resultados muestran que las reglas estáticas de riesgo protegen el capital en mercados bajistas pero sacrifican la participación en las subidas, mientras que la recompensa fraccionaria propuesta logra el comportamiento más robusto entre regímenes, con un ratio de Sharpe estable y caídas máximas contenidas en ambos años de test.

Palabras clave: aprendizaje por refuerzo; optimización de carteras; Proximal Policy Optimization; FinRL; ratio de Sharpe; momentos fraccionarios; GARCH; rentabilidad ajustada al riesgo.

Resumen ejecutivo

1. Motivación y contexto

La gestión de carteras consiste en repartir el capital entre un conjunto de activos con el objetivo de maximizar la rentabilidad ajustada al riesgo. Los enfoques clásicos, herederos de la teoría de carteras de Markowitz, se apoyan en hipótesis estadísticas restrictivas y en reglas estáticas que no se adaptan a los cambios de régimen del mercado. El aprendizaje por refuerzo (RL) ofrece una alternativa atractiva: un agente aprende una política de inversión directamente de los datos, interactuando con un mercado simulado y recibiendo una recompensa por cada decisión, sin necesidad de reglas diseñadas manualmente.

Sin embargo, el RL aplicado a finanzas arrastra un problema fundamental. El agente optimiza la recompensa acumulada y, si esta se define únicamente en términos de retorno, no existe ningún incentivo para controlar el riesgo. El resultado son estrategias agresivas que funcionan bien en mercados alcistas pero amplifican las pérdidas de forma dramática en mercados bajistas. Este trabajo aborda esa cuestión: cómo introducir la conciencia del riesgo en el propio proceso de aprendizaje del agente, y hasta qué punto las distintas formas de hacerlo (reglas duras en el entorno frente a señales de recompensa estadísticamente fundamentadas) condicionan la robustez de la política aprendida ante regímenes de mercado opuestos.

La cuestión no es puramente académica. La medida de riesgo dominante, la desviación estándar que subyace al ratio de Sharpe, presupone momentos finitos de orden cuatro, una hipótesis que las series financieras reales, con colas pesadas y volatilidad agrupada, violan de forma sistemática. Trabajos teóricos recientes de López de Prado et al. demuestran que la inferencia clásica sobre el ratio de Sharpe se invalida precisamente en las condiciones de mercado en las que más se necesita, lo que abre la puerta a medidas alternativas basadas en momentos de orden fraccionario.

2. Metodología

El trabajo se estructura como un pipeline experimental incremental de cinco fases, construido sobre las librerías de código abierto FinRL y Stable-Baselines3. El universo de inversión lo componen diez activos cotizados diversificados por geografía y sector (AAPL, JPM, PG, XOM, ASML, NVO, LVMUY, TSM, TM y BABA). Los agentes se entrenan con datos diarios de 2015 a 2021 y se evalúan fuera de muestra en dos años completos nunca vistos durante el entrenamiento: 2022, un mercado claramente bajista, y 2023, un mercado alcista. Esta separación en dos regímenes opuestos permite medir la robustez de cada estrategia y no solo su rendimiento puntual.

La fase 1 establece la referencia pasiva de compra y mantenimiento (*Buy & Hold*) con pesos iguales. La fase 2 entrena un agente PPO base cuya recompensa es el incremento del valor de la cartera, sin ninguna noción de riesgo. La fase 3 introduce un entorno consciente del riesgo (*Risk-Aware PPO*) que incorpora un mecanismo de stop-loss: cuando el valor de una posición cae por debajo de un umbral respecto a su máximo reciente, el entorno fuerza la liquidación, integrando la gestión del riesgo en la dinámica

del mercado simulado. La fase 4 aplica optimización bayesiana de hiperparámetros con Optuna (40 ensayos) sobre el agente consciente del riesgo, seleccionando la configuración que maximiza el valor de la cartera en validación. La fase 5 constituye la contribución original del trabajo: sustituir el ratio de Sharpe clásico de la señal de recompensa por un ratio de Sharpe de momentos fraccionarios, en el que la desviación estándar del denominador se reemplaza por un momento de orden $p \in (1, 2)$ calculado sobre una ventana móvil de 21 días de negociación. Esta elección, motivada por la teoría estadística reciente, penaliza con más intensidad el riesgo de caída sin ahogar la participación en las subidas, y conserva propiedades inferenciales válidas bajo colas pesadas.

Todas las fases comparten los mismos datos, la misma arquitectura de red y el mismo presupuesto de entrenamiento, de modo que las diferencias de comportamiento sean atribuibles al diseño del entorno y de la recompensa. El código completo del proyecto está disponible públicamente en un repositorio de GitHub para garantizar la reproducibilidad.

3. Resultados

En el mercado bajista de 2022 los resultados confirman el diagnóstico de partida. El agente PPO base, sin control de riesgo, pierde un 29,1 % con una caída máxima del 30,8 %, un resultado peor que el propio mercado (Buy & Hold: −7,5 %). En cambio, el agente consciente del riesgo preserva el capital de forma notable: termina el año bajista en positivo (+2,3 %) con una caída máxima de solo el 2,5 %. La recompensa fraccionaria propuesta cierra prácticamente plana (−0,7 %) con una caída máxima del 3,7 %.

El mercado alcista de 2023 revela la otra cara de las reglas estáticas. El PPO base captura toda la subida (+45,6 %), pero su inconsistencia entre años lo invalida como estrategia: depende por completo del régimen. Las estrategias con stop-loss apenas participan de la subida (+2,9 % el agente consciente del riesgo y +2,4 % la variante optimizada con Optuna), poniendo de manifiesto el coste de oportunidad de las reglas duras. La recompensa de momentos fraccionarios, en cambio, más que duplica esa participación (+5,9 %) y alcanza un ratio de Sharpe de 1,58, prácticamente igual al del mercado (1,61), con una volatilidad anualizada de solo el 3,8 % frente al 14,7 % del índice de referencia.

En conjunto, la estrategia basada en la recompensa fraccionaria es la única que mantiene un ratio de Sharpe estable y caídas máximas contenidas en ambos regímenes: no es la que más gana en un año concreto, pero sí la más robusta entre regímenes, que era precisamente el objetivo de diseño. Un hallazgo adicional relevante es que la optimización de hiperparámetros no compensa un diseño deficiente de la recompensa: la variante optimizada con Optuna no mejora al agente consciente del riesgo de partida, lo que indica que la elección de la señal de recompensa domina sobre el ajuste fino del algoritmo.

4. Conclusiones

Del trabajo se extraen cuatro conclusiones principales. Primera, la conciencia del riesgo no emerge espontáneamente en un agente de RL: debe diseñarse de forma explícita, ya sea en el entorno o en la recompensa. Un agente que solo maximiza retorno es temerario

y estructuralmente inconsistente entre regímenes. Segunda, las reglas estáticas de riesgo, como el stop-loss, son eficaces para preservar capital en mercados bajistas, pero imponen un coste de oportunidad severo en mercados alcistas; esta asimetría constituye un compromiso fundamental documentado experimentalmente en la memoria. Tercera, la recompensa basada en el ratio de Sharpe de momentos fraccionarios logra el mejor equilibrio entre protección y participación, con el comportamiento más robusto entre los dos regímenes de test, y además descansa sobre una base estadística válida en presencia de colas pesadas. Y cuarta, el diseño de la señal de recompensa es la decisión más determinante de todo el sistema, por encima de la optimización de hiperparámetros.

Como líneas futuras se proponen la ampliación del universo a otras clases de activos, la comparación con algoritmos alternativos (SAC, TD3, métodos basados en modelo), la incorporación de costes de transacción y fricciones de mercado realistas, y el aprendizaje adaptativo del orden fraccionario p en función del régimen de mercado vigente.

UNIVERSIDAD PONTIFICIA COMILLAS
Escuela Técnica Superior de Ingeniería (ICAI)
Máster en Ingeniería Industrial

Universidad Pontificia Comillas
ICAI School of Engineering
Master in Industrial Engineering

ANALYSIS AND APPLICATION OF REINFORCEMENT LEARNING TO PORTFOLIO OPTIMIZATION

Author: Jorge González Calvo
Director: Lucía Güitta López
Co-Director: Eduardo César Garrido Merchán

Abstract

This thesis investigates the application of Reinforcement Learning (RL) to equity portfolio management, focusing on the Proximal Policy Optimization (PPO) algorithm within the FinRL framework. Agent behaviour is evaluated across two contrasting market regimes, a bear market (2022) and a bull market (2023), on a universe of ten international assets trained on historical data from 2015 to 2021. The study is organized in five experimental phases: a passive buy-and-hold benchmark, a vanilla PPO baseline, a risk-aware environment with static stop-loss rules, Bayesian hyperparameter optimization via Optuna and, as the original contribution, a reward function based on the fractional-moment Sharpe ratio, motivated by the recent theory of López de Prado et al. (2026) on the breakdown of classical inference under GARCH dynamics and heavy tails. The results show that static risk rules protect capital in bear markets but sacrifice upside participation, whereas the proposed fractional reward achieves the most robust cross-regime behaviour, with a stable Sharpe ratio and contained maximum drawdowns in both test years.

Keywords: Reinforcement Learning; Portfolio Optimization; Proximal Policy Optimization; FinRL; Sharpe Ratio; Fractional Moments; GARCH; Risk-Adjusted Returns.

Executive summary

1. Motivation and context

Portfolio management is the problem of allocating capital across a set of assets so as to maximize risk-adjusted returns. Classical approaches, descending from Markowitz's portfolio theory, rely on restrictive statistical assumptions and static rules that do not adapt to changing market regimes. Reinforcement learning (RL) offers an appealing alternative: an agent learns an investment policy directly from data by interacting with a simulated market and receiving a reward for each decision, with no hand-crafted rules.

RL applied to finance, however, carries a fundamental problem. The agent optimizes cumulative reward and, if the reward is defined purely in terms of returns, there is no incentive whatsoever to control risk. The result is aggressive strategies that perform well in bull markets but dramatically amplify losses in bear markets. This thesis addresses precisely that question: how to introduce risk awareness into the agent's learning process itself, and to what extent the different ways of doing so (hard rules in the environment versus statistically grounded reward signals) condition the robustness of the learned policy across opposing market regimes.

The question is not purely academic. The dominant risk measure, the standard deviation underlying the Sharpe ratio, presupposes finite fourth moments, an assumption that real financial series, with heavy tails and clustered volatility, violate systematically. Recent theoretical work by López de Prado et al. shows that classical inference on the Sharpe ratio breaks down precisely in the market conditions where it is most needed, opening the door to alternative measures based on fractional moments.

2. Methodology

The work is structured as an incremental five-phase experimental pipeline built on the open-source libraries FinRL and Stable-Baselines3. The investment universe consists of ten internationally diversified listed assets (AAPL, JPM, PG, XOM, ASML, NVO, LVMUY, TSM, TM and BABA). Agents are trained on daily data from 2015 to 2021 and evaluated out-of-sample on two full years never seen during training: 2022, a clearly bearish market, and 2023, a bullish one. This two-regime split measures the robustness of each strategy rather than its performance in a single environment.

Phase 1 establishes the passive equal-weight Buy & Hold benchmark. Phase 2 trains a vanilla PPO agent whose reward is the change in portfolio value, with no notion of risk. Phase 3 introduces a risk-aware environment (*Risk-Aware PPO*) incorporating a stop-loss mechanism: when the value of a position falls below a threshold relative to its recent high-water mark, the environment forces liquidation, embedding risk management into the dynamics of the simulated market. Phase 4 applies Bayesian hyperparameter optimization with Optuna (40 trials) to the risk-aware agent, selecting the configuration that maximizes validation portfolio value. Phase 5 constitutes the original contribution of this work: replacing the classical Sharpe ratio in the reward signal with a fractional-moment Sharpe ratio, in which the standard deviation in the denominator is replaced by a moment of order $p \in (1, 2)$ computed over a rolling window of 21 trading days.

This choice, motivated by recent statistical theory, penalizes downside risk more heavily without stifling upside participation, and retains valid inferential properties under heavy tails.

All phases share the same data, network architecture and training budget, so that behavioural differences are attributable to the design of the environment and the reward. The complete project code is publicly available in a GitHub repository to guarantee reproducibility.

3. Results

In the 2022 bear market the results confirm the initial diagnosis. The vanilla PPO agent, with no risk control, loses 29.1% with a maximum drawdown of 30.8%, worse than the market itself (Buy & Hold: -7.5%). The risk-aware agent, by contrast, preserves capital remarkably well: it ends the bear year in positive territory ($+2.3\%$) with a maximum drawdown of only 2.5%. The proposed fractional reward closes essentially flat (-0.7%) with a maximum drawdown of 3.7%.

The 2023 bull market reveals the other side of static rules. The vanilla PPO captures the full rally ($+45.6\%$), but its inconsistency across years invalidates it as a strategy: it depends entirely on the regime. The stop-loss strategies barely participate in the rally ($+2.9\%$ for the risk-aware agent and $+2.4\%$ for the Optuna-optimized variant), exposing the opportunity cost of hard rules. The fractional-moment reward, in contrast, more than doubles that participation ($+5.9\%$) and reaches a Sharpe ratio of 1.58, essentially equal to the market's (1.61), with an annualized volatility of only 3.8% versus 14.7% for the benchmark.

Overall, the fractional-reward strategy is the only one that maintains a stable Sharpe ratio and contained maximum drawdowns in both regimes: it is not the top performer in any single year, but it is the most robust across regimes, which was precisely the design objective. An additional relevant finding is that hyperparameter optimization does not compensate for a deficient reward design: the Optuna-optimized variant does not improve on the risk-aware baseline, indicating that the choice of reward signal dominates over fine-tuning of the algorithm.

4. Conclusions

Four main conclusions follow from this work. First, risk awareness does not emerge spontaneously in an RL agent: it must be designed explicitly, either in the environment or in the reward. An agent that only maximizes returns is reckless and structurally inconsistent across regimes. Second, static risk rules such as stop-losses are effective at preserving capital in bear markets, but impose a severe opportunity cost in bull markets; this asymmetry constitutes a fundamental trade-off documented experimentally in the thesis. Third, the reward based on the fractional-moment Sharpe ratio achieves the best balance between protection and participation, with the most robust behaviour across the two test regimes, while resting on a statistical foundation that remains valid under heavy tails. Fourth, the design of the reward signal is the single most consequential decision in the entire system, above hyperparameter optimization.

Future lines of work include extending the universe to other asset classes, comparing

PPO with alternative algorithms (SAC, TD3, model-based methods), incorporating realistic transaction costs and market frictions, and learning the fractional order p adaptively as a function of the prevailing market regime.

Contents

Acknowledgements **iii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem Statement . . . 2
1.3 Objectives . . . 2
1.4 Scope and Limitations . . . 3
1.5 Document Structure . . . 3

2 State of the Art **5**
2.1 Reinforcement Learning in Portfolio Management . . . 5
2.1.1 Value-Based Methods . . . 6
2.1.2 Policy Gradient Methods . . . 7
2.1.3 Actor-Critic Methods . . . 7
2.2 Proximal Policy Optimization (PPO) . . . 7
2.3 FinRL: An Open-Source Framework for Financial RL . . . 8
2.4 Reward Function Design and Risk-Aware Objectives . . . 8
2.4.1 Volatility-Penalized Rewards . . . 9
2.4.2 Sharpe Ratio-Based Rewards . . . 9
2.4.3 Drawdown Penalties . . . 9
2.4.4 Stop-Loss Mechanisms . . . 9
2.5 Statistical Inference for the Sharpe Ratio . . . 10
2.5.1 Classical Asymptotic Theory . . . 10
2.5.2 Breakdown under GARCH and Heavy Tails . . . 10
2.5.3 Fractional-Moment Sharpe Ratio . . . 11
2.6 Related Work and Research Gaps . . . 11
2.6.1 Deep Reinforcement Learning for Trading and Portfolio Management 11
2.6.2 Research Gaps . . . 12

3 Methodology **13**
3.1 Asset Universe and Data Acquisition . . . 13
3.2 Feature Engineering . . . 14
3.3 Train–Test Split . . . 15
3.4 Trading Environment Configuration . . . 15
3.4.1 Baseline Environment: `StockTradingEnv` . . . 15
3.4.2 Risk-Aware Environment: `StockTradingEnvStopLoss` . . . 16
3.5 State and Action Spaces . . . 16
3.5.1 State Space . . . 16
3.5.2 Action Space . . . 17
3.6 Agent: Proximal Policy Optimization . . . 17
3.7 Hyperparameter Optimization with Optuna . . . 18
3.7.1 Search Space . . . 18

3.7.2 Optimization Protocol . . . 18
3.8 Original Contribution: Fractional-Moment Sharpe Reward . . . 18
3.8.1 Reward Function Formulation . . . 19
3.8.2 Parameter Selection: p . . . 19
3.8.3 Implementation . . . 19
3.9 Out-of-Sample Evaluation Protocol . . . 19

4 Results, Discussion and Comparison 21
4.1 Performance Comparison in the 2022 Bear Market . . . 21
4.1.1 Discussion of the 2022 Results . . . 22
4.2 Performance Comparison in the 2023 Bull Market . . . 23
4.2.1 Discussion of the 2023 Results . . . 24
4.3 Cross-Regime Analysis: The Static Risk Rule Trade-Off . . . 24
4.4 Bayesian Hyperparameter Optimization with Optuna . . . 25
4.4.1 Best Hyperparameters Found . . . 25
4.4.2 Results . . . 26
4.4.3 Discussion: A Negative Result with Positive Implications . . . 26
4.5 Original Contribution: Fractional-Moment Sharpe Reward . . . 27
4.5.1 Results . . . 27
4.5.2 Visual Comparison of All Strategies . . . 27
4.5.3 Drawdown Behavior . . . 29
4.5.4 Aggregate View: Returns and Risk-Adjusted Performance . . . 30
4.6 Final Discussion . . . 31
4.7 Summary of Findings . . . 32

5 Conclusions and Future Work 34
5.1 Summary of the Work . . . 34
5.2 Principal Conclusions . . . 34
5.3 Limitations of the Work . . . 35
5.4 Future Work . . . 36
5.4.1 Methodological Refinements . . . 36
5.4.2 Theoretical Extensions . . . 37
5.4.3 Broader Empirical Validation . . . 37
5.4.4 Practical Deployment Considerations . . . 37
5.5 Closing Remarks . . . 38

6 Alignment with the Sustainable Development Goals 39
6.1 SDG 8: Decent Work and Economic Growth . . . 39
6.2 SDG 9: Industry, Innovation and Infrastructure . . . 39
6.3 SDG 12: Responsible Consumption and Production . . . 39

References 41

List of Figures

2.1 Agent-environment interaction loop 6
2.2 Agent components: model-free vs. model-based 6

3.1 Experimental pipeline 13

4.1 Foundational strategies comparison in 2022 22
4.2 Foundational strategies comparison in 2023 23
4.3 All strategies — 2022 28
4.4 All strategies — 2023 28
4.5 Risk-aware strategies — zoom 2022 29
4.6 Risk-aware strategies — zoom 2023 29
4.7 Drawdown profile — 2022 30
4.8 Drawdown profile — 2023 30
4.9 Total returns by strategy and regime 31
4.10 Annualized Sharpe ratio by strategy and regime 31

List of Tables

3.1 Asset universe 14
3.2 Baseline environment configuration 15
3.3 Risk-Aware environment configuration 16
3.4 PPO hyperparameters 17
3.5 Optuna search space 18

4.1 Performance metrics in the 2022 bear market 22
4.2 Performance metrics in the 2023 bull market 23
4.3 Combined returns across market regimes 24
4.4 Best hyperparameters found by Optuna 25
4.5 Optuna-optimized vs Risk-Aware baseline 26
4.6 Fractional Sharpe results 27

Chapter 1 Introduction

1.1 Motivation

The management of investment portfolios has long been dominated by frameworks rooted in classical financial theory. Since the seminal work of Markowitz [1], mean-variance optimization has provided the conceptual backbone of modern portfolio theory, offering a mathematically elegant trade-off between expected return and risk. Despite its theoretical appeal, this approach rests on assumptions that are frequently violated in practice: static covariance matrices, normally distributed returns, and a static, single-period allocation decision that cannot adapt to changing market conditions.

Financial markets are inherently dynamic, non-stationary environments. Volatility clusters in time [2], return distributions exhibit heavy tails and skewness [3], and regime changes, such as the shift from a prolonged bull cycle to a sharp recessionary downturn, can render any static allocation strategy obsolete within days. A portfolio optimized on historical data under normal conditions will not necessarily survive a period of systemic stress, nor will one calibrated for a bear market efficiently capture the upside of a subsequent recovery.

Reinforcement Learning (RL) offers a fundamentally different paradigm. Rather than solving a one-shot optimization problem, an RL agent learns a *policy*, a mapping from observed market states to allocation decisions, through repeated interaction with a simulated trading environment. The agent is not given the "correct" answer; instead, it explores the space of possible actions and gradually improves its behavior based on a reward signal that reflects financial performance. This framework is naturally suited to sequential decision-making under uncertainty, making it a compelling alternative to static methods for portfolio management.

The growing availability of open-source frameworks such as FinRL [4] and deep learning libraries such as Stable-Baselines3 [5] has substantially lowered the barrier to applying state-of-the-art RL algorithms, including Proximal Policy Optimization (PPO) [6], to realistic financial datasets. Nevertheless, the performance of these agents is critically sensitive to how the learning objective is defined. A reward function based solely on raw portfolio returns may produce agents that take excessive risk or fail to distinguish between profitable volatility and catastrophic drawdown.

This thesis is motivated by the recognition that the design of the reward function is the central unsolved problem in RL-based portfolio management. In particular, the Sharpe ratio, the canonical risk-adjusted performance measure, may itself be a flawed basis for reward design when returns exhibit conditional heteroskedasticity and heavy tails, as demonstrated theoretically by López de Prado et al. [7]. Addressing this limitation constitutes the original scientific contribution of the present work.

1.2 Problem Statement

This thesis investigates a central tension in algorithmic portfolio management: the *trade-off between downside protection and upside capture across different market regimes.*

The study considers two contrasting market periods as out-of-sample test environments:

- *2022 (Bear market):* A year characterized by sharply falling equity prices, driven by rising interest rates, geopolitical uncertainty, and the unwinding of post-pandemic valuations. The passive buy-and-hold benchmark returned -7.18%.
- *2023 (Bull market):* A year of strong recovery, particularly in the technology sector, with the benchmark returning $+24.86\%$.

Experimental results obtained during this project reveal that static risk-control mechanisms, specifically, hard stop-loss rules and minimum cash reserve constraints applied to the RL environment, produce highly asymmetric behavior across these regimes. While such constraints successfully protect capital in bear markets (reducing losses from -4.20% to -0.77%), they simultaneously act as a constraint on upside participation in bull markets (capping returns at $+0.48\%$ against a $+24.86\%$ benchmark). This phenomenon is referred to throughout this thesis as the *static risk rule trade-off.*

At a deeper level, the problem is one of risk measurement. The classical Sharpe ratio penalizes all volatility symmetrically, treating upward price movements as equally undesirable as downward ones. Under standard Gaussian assumptions, this is harmless; under GARCH dynamics with infinite fourth moments, a regime that empirical evidence suggests is common in financial markets, the standard Sharpe estimator becomes statistically unstable, and inference based on it breaks down [7].

The scientific problem addressed in this thesis is therefore:

> *Can a reinforcement learning agent, trained with a reward function based on a statistically robust alternative to the classical Sharpe ratio, achieve better risk-adjusted performance simultaneously across bear and bull market regimes, overcoming the limitations of static risk rules?*

1.3 Objectives

The work is organized around three main objectives:

1. *Baseline reproduction and evaluation.* Implement and validate the default FinRL PPO agent on a portfolio of ten assets, establishing a rigorous empirical baseline against passive investment strategies and evaluating its behavior across the 2022 bear and 2023 bull markets.
2. *Risk-aware environment design and hyperparameter optimization.* Design a custom trading environment incorporating static risk-control mechanisms (stop-loss, minimum cash reserve, profit/loss ratio), analyze the resulting performance trade-off, and apply Bayesian hyperparameter optimization via Optuna [8] to partially recover upside performance without sacrificing downside protection.

3. *Original contribution: fractional-moment Sharpe ratio.* Implement a novel reward function based on the fractional-moment signal-to-noise functional proposed by López de Prado et al. [7], which remains statistically valid under GARCH returns and heavy-tailed distributions. The goal is to obtain an agent that dynamically adapts its risk behavior across market regimes, surpassing the static rule approach in bull markets while maintaining comparable protection in bear markets.

1.4 Scope and Limitations

The following constraints define the boundaries of this study:

- *Asset universe:* Ten large-cap equities selected from diverse geographies and indices, covering sectors including technology, healthcare, financials, energy, and consumer goods. The specific tickers and their rationale are described in Chapter 3.
- *Data period:* Historical daily price data from January 2015 to December 2023, sourced via Yahoo Finance through FinRL's `YahooDownloader`. The training window spans 2015–2021; the test window covers 2022 and 2023 separately.
- *Algorithm:* All experiments use PPO as implemented in Stable-Baselines3, accessed through the FinRL interface. Alternative algorithms (DDPG, SAC, A2C) are discussed in the state of the art but not experimentally evaluated.
- *Transaction costs:* The FinRL trading environment used in this study applies a fixed transaction cost rate per trade. Market impact, slippage, and bid-ask spreads are not modeled explicitly.
- *Long-only constraint:* The agent may only take long positions or hold cash; short selling is not permitted.
- *Generalization:* Results are specific to the asset universe, time period, and market conditions studied. The findings should not be interpreted as a general claim about the performance of RL in all financial markets.

1.5 Document Structure

The remainder of this thesis is organized as follows:

Chapter 2, State of the Art reviews the relevant literature on reinforcement learning for portfolio management, covering the main RL algorithms applied in financial contexts, reward function design, risk-aware objectives, and the statistical theory of Sharpe ratio inference under non-Gaussian return distributions.

Chapter 3, Methodology describes the complete experimental pipeline: asset selection and data preprocessing, the FinRL environment configuration, the PPO agent architecture, the risk-aware environment design, the Bayesian optimization procedure, and the mathematical formulation of the proposed fractional-moment reward function.

Chapter 4, Results, Discussion and Comparison presents the empirical results of all experimental phases, provides a quantitative comparison against benchmarks, and analyzes the static risk rule trade-off in depth.

Chapter 5, Conclusions and Future Work summarizes the principal findings of the thesis, discusses their implications for practical portfolio management, and outlines directions for future research.

Chapter 2 State of the Art

2.1 Reinforcement Learning in Portfolio Management

Over the past decade, reinforcement learning has emerged as a compelling alternative to classical portfolio optimization methods. Early works applied tabular or shallow RL algorithms to simplified allocation problems with discrete action spaces. The deep learning revolution enabled a qualitative leap: deep neural networks can now model the complex, non-linear relationships between market state observations and continuous portfolio weight decisions, opening the door to practical deployment on multi-asset portfolios with realistic constraints. Seminal contributions such as the direct reinforcement approach of Moody and Saffell [9] and its deep extension by Deng et al. [10] demonstrated early on that an agent could learn profitable trading signals end-to-end, without an intermediate forecasting stage.

The standard RL formulation for portfolio management treats the allocation problem as a Markov Decision Process (MDP) [11, 12]. At each discrete time step t, the agent observes a state $s_t \in \mathcal{S}$ encoding current market information, typically price histories, technical indicators, and portfolio positions, and selects an action $a_t \in \mathcal{A}$ representing the new target portfolio weights. The environment transitions to a new state s_{t+1} according to the market dynamics, and the agent receives a scalar reward r_t reflecting the financial outcome of its decision. The agent's objective is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t \right], \tag{2.1}$$

where $\gamma \in [0, 1)$ is a discount factor that governs the agent's preference for immediate versus future rewards.

Figure 2.1 illustrates this interaction loop. At each step t, the agent receives the current state s_t and reward r_t, selects action a_t, and the environment transitions to a new state s_{t+1} with associated reward r_{t+1}.

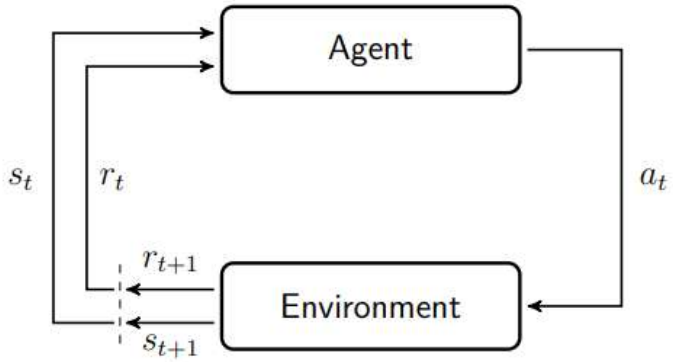

Figure 2.1: The agent-environment interaction loop in reinforcement learning. At each time step t, the agent observes state s_t and reward r_t, selects action a_t, and the environment returns the next state s_{t+1} and reward r_{t+1}.

The literature distinguishes three main families of deep RL algorithms, each with different suitability for portfolio management:

Figure 2.2 illustrates the relationship between these components. Model-free methods work directly with experience to learn a value function or policy, while model-based methods additionally build an explicit model of the environment for planning.

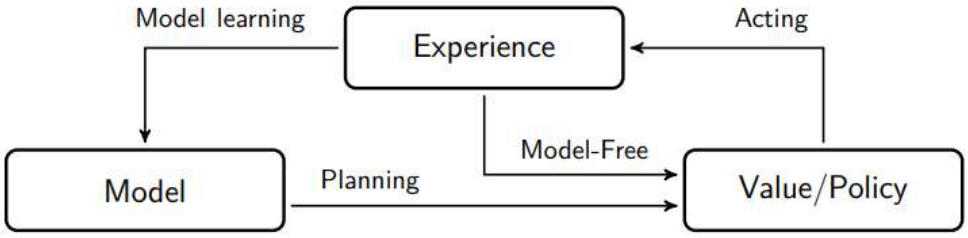

Figure 2.2: Overview of the components of a reinforcement learning agent and their relationships. Model-free methods use experience to directly update value functions and policies. Model-based methods additionally build a model of the environment and use planning to derive optimal behavior.

2.1.1 Value-Based Methods

Value-based methods, typified by Deep Q-Networks (DQN) [13] and rooted in the classical Q-learning algorithm [14], learn to estimate the expected return from each state-action pair. They are well-suited to problems with discrete, finite action spaces. In portfolio management, this restricts their application to simplified settings such as binary long/flat decisions on individual assets or discretized leverage levels. The curse of dimensionality makes pure value-based methods impractical for continuous weight allocation across $N > 2$ assets.

2.1.2 Policy Gradient Methods

Policy gradient methods [15] directly parameterize and optimize the policy π_θ with respect to the performance objective $J(\pi_\theta)$. The gradient of the objective is estimated from sampled trajectories:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) \cdot G_t]\,, \tag{2.2}$$

where G_t is the return from step t onward. These methods naturally handle continuous action spaces and can represent stochastic policies, providing built-in exploration. However, naïve policy gradient estimates suffer from high variance, which can make training unstable on financial data with heavy-tailed returns.

2.1.3 Actor-Critic Methods

Actor-critic architectures combine a policy network (the *actor*) with a value function estimator (the *critic*). The critic provides a baseline or advantage estimate that reduces gradient variance while preserving unbiasedness. Algorithms such as A2C [16], DDPG [17], TD3 [18], and SAC [19] belong to this family and have all been applied to portfolio management tasks with varying degrees of success.

2.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization [6] has become one of the most widely adopted RL algorithms in both academic research and practical applications. It belongs to the actor-critic family and addresses the instability of earlier policy gradient methods through a clipped surrogate objective that constrains the magnitude of policy updates at each training step. It can be understood as a first-order, more easily implemented alternative to Trust Region Policy Optimization [20], which enforces a comparable constraint through a computationally heavier trust-region formulation.

The core innovation of PPO is the clipped objective function:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)\,\hat{A}_t,\ \text{clip}(r_t(\theta),\,1-\varepsilon,\,1+\varepsilon)\,\hat{A}_t\Big)\Big]\,, \tag{2.3}$$

where $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability ratio between the updated and old policies, $\hat{A}_t$ is the estimated advantage at step t, typically obtained through generalized advantage estimation [21], and ε (typically 0.1–0.2) is the clipping parameter. The min operator ensures that the update is conservative: the surrogate loss is only improved when the policy moves in the advantageous direction without deviating excessively from the previous iterate.

The full PPO loss combines the clipped policy objective, a value function regression term, and an entropy bonus that encourages exploration:

$$L(\theta) = L^{\text{CLIP}}(\theta) - c_1\,L^{\text{VF}}(\theta) + c_2\,S[\pi_\theta](s_t), \tag{2.4}$$

where c_1 and c_2 are weighting coefficients and $S[\pi_\theta](s_t)$ is the entropy of the policy at state s_t.

In the context of portfolio management, the stability properties of PPO are particularly valuable. Financial time series exhibit non-stationarity, volatility clustering, and occasional extreme observations that can produce large gradient updates. The clipping mechanism of PPO provides a natural safeguard against these destabilizing events. Several studies have reported that PPO outperforms or matches more complex algorithms such as DDPG and SAC on portfolio benchmarks, while being significantly easier to tune [4].

2.3 FinRL: An Open-Source Framework for Financial RL

FinRL [4] is an open-source Python library specifically designed for applying deep reinforcement learning to financial problems. It provides an end-to-end pipeline integrating:

- *Data acquisition:* Wrappers for Yahoo Finance, Alpaca, and other market data providers via a unified `YahooDownloader` interface.
- *Feature engineering:* Automated computation of technical indicators (MACD, RSI, Bollinger Bands, VIX) through the `FeatureEngineer` module.
- *Trading environments:* OpenAI Gym-compatible `StockTradingEnv` implementations supporting multi-asset continuous allocation, configurable reward functions, and optional risk constraints.
- *Agent integration:* Seamless wrappers for Stable-Baselines3 [5] algorithms including PPO, A2C, DDPG, TD3, and SAC.

The standard FinRL reward function at step t is the change in portfolio value:

$$r_t = v_{t+1} - v_t, \tag{2.5}$$

where v_t denotes the total portfolio value at time t. This formulation encourages the agent to maximize absolute wealth accumulation but provides no explicit incentive for risk control, which motivates the modifications explored in this thesis.

FinRL has been adopted as a benchmark platform in a growing body of academic work. Its standardized environments enable direct comparison of algorithm variants across studies, making it well-suited for the systematic evaluation carried out in Chapters 3 and 4.

2.4 Reward Function Design and Risk-Aware Objectives

The reward function is arguably the most consequential design choice in an RL-based trading system. It encodes what the agent is trained to optimize and therefore determines, more than any other component, whether the learned policy will be useful in practice.

The simplest reward, raw portfolio return at each step, encourages wealth maximization but ignores the path taken to achieve it. An agent optimizing this objective may learn highly volatile strategies that occasionally produce large gains but suffer catastrophic losses in adverse conditions. A substantial body of literature has therefore explored risk-adjusted reward formulations, drawing on the broader theory of coherent risk measures [22].

2.4.1 Volatility-Penalized Rewards

A common approach is to penalize the variance or standard deviation of returns within a time window:

$$r_t = \bar{R}_t - \lambda\,\sigma_t, \tag{2.6}$$

where $\bar{R}_t$ is the mean return over a recent window, σ_t is its standard deviation, and λ is a risk-aversion parameter. This directly encodes the mean-variance trade-off in the reward signal.

2.4.2 Sharpe Ratio-Based Rewards

The episodic or rolling Sharpe ratio is another widely used reward proxy. At the episodic level:

$$r_{\text{episode}} = \frac{\bar{R}}{\hat{\sigma}}, \tag{2.7}$$

where the mean and standard deviation are computed over the full episode. Online or step-level Sharpe approximations are also used, where running statistics are maintained and updated at each step.

2.4.3 Drawdown Penalties

Maximum drawdown, the largest peak-to-trough decline over a period, is a widely used risk metric in practice. Incorporating it into the reward:

$$r_t = R_t - \mu_{\text{DD}} \cdot \text{DD}_t, \tag{2.8}$$

where DD_t is the drawdown at time t and μ_{DD} is a penalty coefficient. Drawdown-penalized rewards tend to produce agents that are more cautious about entering loss-making positions. Closely related tail-risk measures, most notably Value-at-Risk and Conditional Value-at-Risk (expected shortfall) [23], have likewise been incorporated into reward signals to target the left tail of the return distribution explicitly.

2.4.4 Stop-Loss Mechanisms

Hard stop-loss rules can be implemented directly within the trading environment rather than the reward function. When the portfolio value drops below a predefined threshold relative to a recent high-water mark, the environment forces liquidation to cash, preventing further losses. While effective at cutting losses in bear markets, such rules can prevent the agent from re-entering the market promptly, resulting in missed recovery gains, the central trade-off identified in this thesis.

2.5 Statistical Inference for the Sharpe Ratio

The Sharpe ratio [24] is defined as the expected excess return of a strategy divided by its standard deviation:

$$\mathrm{SR} = \frac{\mu}{\sigma}, \tag{2.9}$$

where $\mu = \mathbb{E}[R_t]$ and $\sigma = \sqrt{\mathrm{Var}(R_t)}$. In practice, both quantities must be estimated from finite samples, yielding the plug-in estimator:

$$\widehat{\mathrm{SR}}_n = \frac{\bar{R}_n}{\hat{\sigma}_n}. \tag{2.10}$$

2.5.1 Classical Asymptotic Theory

Lo [25] established the foundational asymptotic framework for Sharpe ratio inference. Under the assumption that the joint central limit theorem holds for $(\bar{R}_n, \overline{R_n^2})$, the plug-in estimator satisfies:

$$\sqrt{n}\,(\widehat{\mathrm{SR}}_n - \mathrm{SR}) \xrightarrow{d} \mathcal{N}(0, V), \tag{2.11}$$

where V depends on the long-run covariance structure of returns. Under serial dependence and non-normality, Lo's approximation incorporates corrections for autocorrelation, skewness γ_3, and excess kurtosis γ_4:

$$\widehat{\mathrm{SR}}_n \overset{a}{\sim} \mathcal{N}\left(\mathrm{SR},\ \frac{1}{T}\left[\frac{1+\rho}{1-\rho} - \frac{1+\rho+\rho^2}{1-\rho^2}\,\gamma_3\,\mathrm{SR} + \frac{1+\rho^2}{1-\rho^2}\,\frac{\gamma_4 - 1}{4}\,\mathrm{SR}^2\right]\right), \tag{2.12}$$

where ρ is the first-order autocorrelation of returns.

2.5.2 Breakdown under GARCH and Heavy Tails

A fundamental limitation of the classical framework is its reliance on the existence of a finite fourth moment $\mathbb{E}[R_t^4] < \infty$. This assumption sits uneasily with the long-documented stylized facts of financial returns, which exhibit heavy tails, volatility clustering, and leptokurtosis [26, 27]. López de Prado et al. [7] prove a structural necessity theorem: if the plug-in Sharpe estimator satisfies $\sqrt{n}$-regularity, then the empirical second moment of returns must also satisfy $\sqrt{n}$-regularity, which in turn requires a finite fourth moment.

In heavy-tailed regimes typical of GARCH processes with tail index $\kappa \in (2, 4)$, the fourth moment is infinite. Partial sums of squared returns converge to stable (non-Gaussian) limits at rate $n^{1/\alpha}$ for $\alpha = \kappa/2 \in (1, 2)$, not at the $\sqrt{n}$ rate. This means the classical Gaussian inference procedure for the Sharpe ratio is invalid precisely in the market conditions, high volatility, extreme events, where it is most commonly applied.

Under a GARCH(1,1) model with parameters α and β ($\phi = \alpha + \beta < 1$), López de Prado et al. derive a closed-form asymptotic variance:

$$V_{\mathrm{GARCH}} = 1 + \frac{\mu^2}{4\sigma^2} \cdot \frac{(1-\beta)^2\,(\kappa_z - 1)\,(1+\phi)}{(1-\phi)\left(1 - \alpha^2\kappa_z - 2\alpha\beta - \beta^2\right)}, \tag{2.13}$$

which makes explicit how volatility clustering (through ϕ) and leptokurtosis (through $\kappa_z = \mathbb{E}[z_t^4]$) inflate the statistical uncertainty of the Sharpe estimator.

2.5.3 Fractional-Moment Sharpe Ratio

To address the breakdown of classical inference in heavy-tailed regimes, López de Prado et al. [7] introduce an alternative signal-to-noise functional based on fractional moments. For $p \in (1, 2)$, define the fractional-moment scale estimator:

$$\hat{\sigma}_{p,n} = \left(\frac{1}{n} \sum_{t=1}^{n} |R_t - \bar{R}_n|^p \right)^{1/p}, \tag{2.14}$$

and the corresponding fractional Sharpe estimator:

$$\widehat{\mathrm{SR}}_{p,n} = \frac{\bar{R}_n}{\hat{\sigma}_{p,n}}. \tag{2.15}$$

The key property of this estimator is that it requires only $\mathbb{E}|R_t|^{p+\delta} < \infty$ for some $\delta > 0$, which holds whenever $p < \kappa$. Since $\kappa > 2$ in stationary GARCH processes (by assumption), choosing $p \in (1, 2)$ guarantees that $\widehat{\mathrm{SR}}_{p,n}$ admits a standard $\sqrt{n}$-Gaussian central limit theorem even when the classical estimator does not.

From the perspective of reward design, $\widehat{\mathrm{SR}}_{p,n}$ has an important financial interpretation: by using a sub-quadratic norm in the denominator, it penalizes volatility *less aggressively* than the classical estimator for large deviations, while still being sensitive to small fluctuations. This asymmetry means that an agent rewarded with $\widehat{\mathrm{SR}}_{p,n}$ will be less averse to the "good" volatility associated with strong bull-market returns, while still penalizing the "bad" volatility associated with heavy-tailed drawdowns. The theoretical motivation for using this functional as a reward signal in the PPO agent constitutes the original contribution of this thesis (see Chapter 3).

2.6 Related Work and Research Gaps

2.6.1 Deep Reinforcement Learning for Trading and Portfolio Management

The application of deep reinforcement learning to trading has evolved rapidly over the last two decades, and it is instructive to situate the present work within that trajectory. The earliest direct-reinforcement approaches [9] optimized a differentiable performance measure, an online Sharpe ratio, with respect to the parameters of a recurrent trading system, already recognizing that the choice of objective is central to the behaviour of the resulting strategy. Deng et al. [10] extended this idea with deep and fuzzy representations, but both works were limited to single-asset or small-scale settings.

The transition to genuine multi-asset portfolio management was driven by deep architectures tailored to the allocation problem. Jiang et al. [28] introduced the Ensemble of Identical Independent Evaluators framework, in which a convolutional network maps the price tensor of each asset to a portfolio weight, trained directly on the logarithmic return of the portfolio. Yang et al. [29] proposed an ensemble that dynamically selects among PPO, A2C, and DDPG according to recent risk-adjusted performance, reporting improved robustness over any single agent. These works established deep RL as competitive with classical baselines, but they optimize raw or log returns and address

risk only indirectly, through the choice of algorithm rather than the design of the reward.

The consolidation of the field around shared infrastructure came with the FinRL ecosystem [4, 30] and its benchmark extension FinRL-Meta [31], which standardize data pipelines, trading environments, and agent interfaces. This thesis builds directly on that ecosystem, which makes its experiments reproducible and its results directly comparable to the broader literature. Complementary studies have explored model-free deep RL for trading with explicit volatility targeting [32] and the closely related problem of hedging derivative portfolios under transaction costs through deep learning [33]. Hyperparameter selection in these pipelines, historically performed by hand, has increasingly been delegated to principled Bayesian optimization [34, 35], a strategy adopted in this work through the Optuna framework.

Against this backdrop, the contribution of the present thesis is specific and complementary. Rather than proposing a new architecture or ensemble, it isolates the reward signal as the object of study and grounds its design in the formal statistical theory of the Sharpe ratio under heavy tails [7]. Where prior work typically rewards raw or log returns [28, 29], or penalizes variance under an implicit Gaussian assumption, this work introduces a fractional-moment objective that remains statistically well-behaved precisely in the heavy-tailed regimes where conventional risk-adjusted rewards lose their theoretical justification.

2.6.2 Research Gaps

A growing body of empirical literature has confirmed that PPO and related actor-critic methods can outperform passive benchmarks in controlled backtests, particularly when combined with technical indicators and risk constraints [4, 29]. However, several gaps remain.

First, there is no consensus on the optimal reward function for portfolio management. Many published results rely on ad hoc formulations with limited theoretical justification, making it difficult to interpret differences in performance across studies.

Second, while static risk rules such as stop-losses are effective in bear markets, their adverse effect on bull-market performance, the trade-off documented experimentally in this thesis, has received limited systematic attention in the RL literature.

Third, the statistical properties of risk-adjusted reward signals have rarely been analyzed formally. The assumption that variance is a valid risk measure for all market conditions underlies most existing reward designs, yet this assumption is violated precisely when robust risk management is most critical.

The present work addresses these gaps by providing empirical evidence of the static risk rule trade-off, by grounding the reward function design in the formal statistical theory of [7], and by proposing a reward signal with provably better statistical properties under realistic market dynamics.

Chapter 3 Methodology

This chapter describes the complete experimental pipeline used in this thesis: from raw market data to a fully trained, evaluated reinforcement learning agent. The pipeline is illustrated in Figure 3.1 and proceeds through five sequential stages: (i) asset selection and data acquisition, (ii) preprocessing and feature engineering, (iii) train–test splitting, (iv) environment configuration and agent training, and (v) out-of-sample evaluation. Each stage is described in detail in the sections that follow.

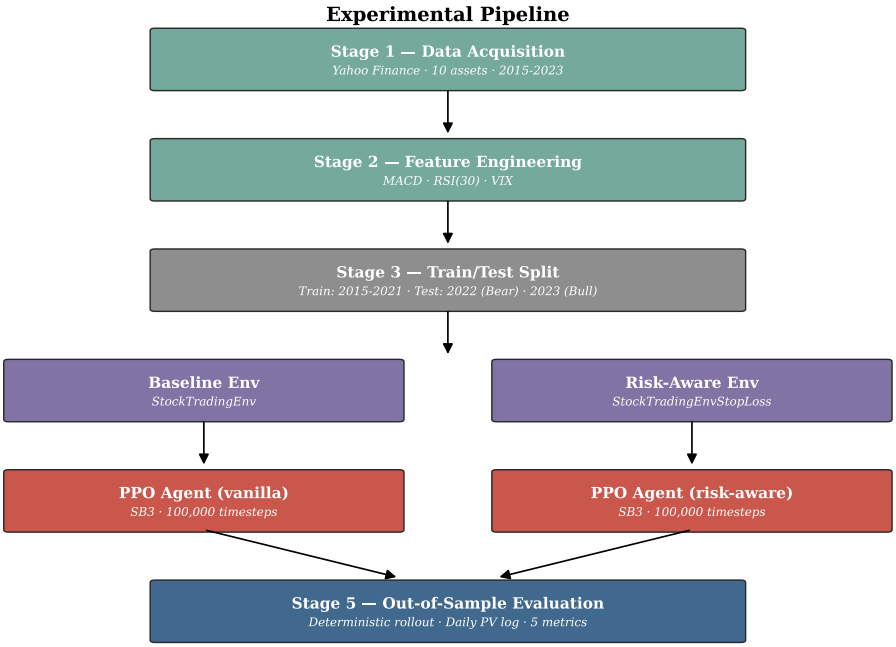

Figure 3.1: Overall experimental pipeline for the RL-based portfolio optimization system. Data flows from raw Yahoo Finance prices through feature engineering, into the trading environment, where the PPO agent interacts iteratively to learn an allocation policy. The trained model is then evaluated out-of-sample on the 2022 and 2023 test windows.

3.1 Asset Universe and Data Acquisition

The investment universe consists of ten internationally diversified large-cap equities, selected to provide exposure across sectors and geographies while maintaining sufficient liquidity for daily trading. Table 3.1 lists the tickers used in this study together with their sector, country of primary listing, and the benchmark market index to which each asset belongs.

Table 3.1: Ten-asset portfolio universe used throughout the experiments, together with the reference market index of each asset's primary listing.

Ticker	Company	Sector	Country	Index
AAPL	Apple Inc.	Technology	USA	S&P 500
JPM	JPMorgan Chase & Co.	Financials	USA	S&P 500
PG	Procter & Gamble Co.	Consumer Staples	USA	S&P 500
XOM	Exxon Mobil Corp.	Energy	USA	S&P 500
ASML	ASML Holding N.V.	Technology	Netherlands	AEX
NVO	Novo Nordisk A/S	Healthcare	Denmark	OMX C25
LVMUY	LVMH Moët Hennessy Louis Vuitton	Consumer Discretionary	France	CAC 40
TSM	Taiwan Semiconductor Manufacturing	Technology	Taiwan	TAIEX
TM	Toyota Motor Corp.	Consumer Discretionary	Japan	Nikkei 225
BABA	Alibaba Group Holding Ltd.	Consumer Discretionary	China	Hang Seng

Daily Open-High-Low-Close-Volume (OHLCV) data was retrieved through the `YahooDownloader` class provided by FinRL, which interfaces with the Yahoo Finance public API. The full data range covers the period from *January 1, 2015 to December 31, 2023*, providing approximately nine years of historical observations across the ten tickers.

3.2 Feature Engineering

The raw price series alone do not constitute a sufficient state representation for an RL agent. Technical indicators provide additional context about price momentum, overbought/oversold conditions, and broader market volatility. The `FeatureEngineer` module of FinRL was used to compute the following indicators for each ticker on each trading day:

- *MACD (Moving Average Convergence Divergence):* A trend- following momentum indicator computed as the difference between the 12-day and 26-day exponential moving averages of the closing price. Positive values indicate upward momentum; negative values indicate downward momentum.
- *RSI (Relative Strength Index, 30-day window):* A bounded oscillator $\in [0, 100]$ measuring the magnitude of recent price changes. Values above 70 typically indicate overbought conditions; values below 30 indicate oversold conditions.
- *VIX (Volatility Index):* The CBOE Volatility Index, representing the market's expectation of 30-day forward-looking volatility implied by S&P 500 options. Unlike MACD and RSI, VIX is a market-wide variable rather than ticker-specific, but it is broadcast to all assets to inform the agent about systemic risk conditions.

After feature engineering, each daily observation for each ticker contains the original OHLCV fields plus the three derived indicators, yielding the state representation described in Section 3.5.

3.3 Train–Test Split

The dataset is partitioned into a single training window and two disjoint out-of-sample test windows, selected to capture opposite market regimes:

- *Training set:* January 1, 2015 – December 31, 2021 (∼1760 trading days). This window contains diverse market conditions including the late stages of the post-2008 recovery, the 2018 correction, the 2020 COVID-19 crash, and the subsequent rebound, providing the agent with broad exposure to varied dynamics.
- *Test set 2022 (bear market):* January 1, 2022 – December 31, 2022 (251 trading days). A year of pronounced equity declines driven by accelerating monetary tightening and geopolitical shocks.
- *Test set 2023 (bull market):* January 1, 2023 – December 31, 2023 (249 trading days). A year of strong recovery, led by the technology sector.

To ensure compatibility with FinRL's indexing logic, which expects sequential integer indices aligned with unique trading days, each partition is sorted by date and ticker and re-indexed using `pandas.factorize()` on the date column.

3.4 Trading Environment Configuration

Two distinct environments are used in this study, corresponding to two of the experimental phases described in Chapter 4.

3.4.1 Baseline Environment: `StockTradingEnv`

The default FinRL trading environment serves as the baseline. It exposes no explicit risk-management mechanisms and rewards the agent purely on the change in total portfolio value at each step. The configuration is summarized in Table 3.2.

Table 3.2: Configuration parameters for the baseline FinRL environment.

Parameter	**Value**
Initial capital	100 000 USD
Maximum shares per trade ($h_{\max}$)	100
Buy transaction cost	0.1 %
Sell transaction cost	0.1 %
Reward scaling factor	10^{-4}
Technical indicators	MACD, RSI(30), VIX
Initial share holdings	0 (all cash)

3.4.2 Risk-Aware Environment: `StockTradingEnvStopLoss`

The risk-aware environment extends the baseline with three additional mechanisms designed to encourage capital preservation. The configuration is summarized in Table 3.3.

Table 3.3: Additional risk-control parameters in the `StockTradingEnvStopLoss` environment.

Parameter	Value
`stoploss_penalty`	0.9 (forced sale on 10 % drop)
`profit_loss_ratio`	2
`cash_penalty_proportion`	0.10
`patient` mode	`True`
`random_start`	`False` (training: `True`)
Daily information columns	close, volume, MACD, RSI(30), VIX

The semantic meaning of each risk-control parameter is as follows:

- *Stop-loss penalty (0.9):* If the price of any held asset drops to 90% or less of its purchase price (i.e., a 10% loss), the environment forcibly liquidates the position. This caps individual- position losses at 10%.
- *Profit/loss ratio (2):* The reward function asymmetrically penalizes losing trades twice as much as it rewards winning trades of equal magnitude, discouraging high-variance "gambling" behavior.
- *Cash penalty proportion (0.10):* The agent is penalized whenever its cash balance falls below 10% of total portfolio value, enforcing a minimum liquidity buffer.
- *Patient mode:* When the agent cannot execute a buy order due to insufficient cash, the environment waits rather than forcing a partial fill or skipping the action.

3.5 State and Action Spaces

3.5.1 State Space

The state vector at time t, denoted $s_t \in \mathbb{R}^{|\mathcal{S}|}$, encodes all information available to the agent for decision-making. For the baseline environment with $N = 10$ assets and 3 technical indicators, the state dimensionality is computed as:

$$|\mathcal{S}| = 1 + 2N + I \cdot N = 1 + 2(10) + 3(10) = 51, \tag{3.1}$$

where the components are:

- 1 scalar for the current cash balance,
- N values for the closing price of each asset,

- N values for the number of shares currently held in each asset,
- $I \cdot N$ values for the technical indicators across all assets ($I = 3$ indicators).

The risk-aware environment uses a slightly different state encoding via the via the `daily_information_cols` parameter, which includes close, volume, MACD, RSI(30), and VIX directly, producing a comparable but not identical representation.

3.5.2 Action Space

The action space is continuous and of dimension $N = 10$:

$$a_t \in [-1, 1]^{10}, \tag{3.2}$$

where the sign of each component indicates the direction (buy or sell) and the magnitude is scaled to a maximum of $h_{\max} = 100$ shares per transaction. The agent therefore decides simultaneously, on each trading day, how many shares to buy or sell across each of the ten assets.

3.6 Agent: Proximal Policy Optimization

The PPO algorithm was selected for the reasons discussed in Section 2.2: stability under non-stationary financial data, support for continuous action spaces, and well-validated implementation in the Stable-Baselines3 library. The default architecture, an MLP with two hidden layers of 64 units each and `tanh` activations, was used for both the actor (policy) and the critic (value) networks.

Training hyperparameters for the baseline configurations are listed in Table 3.4. These default values are subsequently optimized via Optuna in Section 3.7.

Table 3.4: Default PPO hyperparameters used in the baseline and Risk-Aware experimental phases.

Hyperparameter	**Value**
Learning rate	3×10^{-4}
Number of steps per update (n_{steps})	2048
Batch size	64
Discount factor γ	0.99
GAE coefficient λ	0.95
PPO clipping range ε	0.2
Entropy coefficient c_2	0.0
Value function coefficient c_1	0.5
Total training timesteps	100 000
Policy network architecture	MLP $[64, 64]$, `tanh`

3.7 Hyperparameter Optimization with Optuna

To investigate whether the trade-off identified in Chapter 4 admits a better operating point through hyperparameter tuning alone, a Bayesian optimization procedure is implemented using the Optuna framework [8].

3.7.1 Search Space

The optimization searches over the following five PPO hyperparameters, which are most likely to influence the agent's risk-taking behavior:

Table 3.5: Hyperparameter search space used in the Bayesian optimization stage.

Parameter	Range / Choices	Sampling
Learning rate	$[10^{-5}, 10^{-3}]$	log-uniform
n_{steps}	$\{1024, 2048, 4096\}$	categorical
Batch size	$\{64, 128, 256\}$	categorical
Discount factor γ	$[0.95, 0.9999]$	uniform
Entropy coefficient	$[10^{-5}, 10^{-1}]$	log-uniform

3.7.2 Optimization Protocol

The optimization is performed using the Tree-structured Parzen Estimator (TPE) sampler. Each trial trains a fresh PPO agent on the 2015–2021 training set for 50 000 timesteps (reduced from 100 000 to control wall-clock time), and evaluates the resulting policy on the 2023 validation set. The objective function returns the final portfolio value on the validation set, which Optuna maximizes.

A total of 20 trials are conducted. Optimization progress is persisted in an SQLite database, enabling automatic resumption in the event of system interruption.

The best hyperparameters obtained, together with the resulting agent's performance, are reported in Section 4.4 of Chapter 4.

3.8 Original Contribution: Fractional-Moment Sharpe Reward

The fourth and final experimental phase introduces the original contribution of this thesis: the replacement of the variance-based risk measure embedded in the standard Sharpe-ratio reward with a robust, fractional-moment alternative motivated by the theoretical results of [7] reviewed in Section 2.5.3.

3.8.1 Reward Function Formulation

Let $\{R_t\}_{t=1}^{n}$ denote the sequence of daily portfolio returns observed within an episode (or rolling window). The standard window-Sharpe reward is:

$$r_n^{\mathrm{SR}} = \frac{\bar{R}_n}{\hat{\sigma}_n}, \quad \hat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(R_t - \bar{R}_n\right)^2}. \tag{3.3}$$

The proposed fractional-moment Sharpe reward replaces $\hat{\sigma}_n$ with the fractional scale estimator $\hat{\sigma}_{p,n}$ defined in Equation (2.14):

$$r_n^{\mathrm{FSR}} = \frac{\bar{R}_n}{\hat{\sigma}_{p,n}}, \quad \hat{\sigma}_{p,n} = \left(\frac{1}{n}\sum_{t=1}^{n}\left|R_t - \bar{R}_n\right|^p\right)^{1/p}, \quad p \in (1,2). \tag{3.4}$$

3.8.2 Parameter Selection: p

The fractional moment exponent p controls the trade-off between sensitivity to small fluctuations ($p \to 2$, recovering classical variance) and robustness to heavy-tailed extremes ($p \to 1$, approaching the mean absolute deviation). Following the analysis in [7], the valid range is $p \in (1, \kappa)$, where κ is the tail index of the return distribution. For typical equity returns, $\kappa \in (2, 4)$, so values of p in the range $[1.3, 1.7]$ are theoretically justified.

In this work, $p = 1.5$ is adopted as a balanced default, corresponding to the geometric midpoint of the admissible range. A sensitivity analysis across $p \in \{1.3, 1.5, 1.7\}$ is performed and reported in Section 4.5.

3.8.3 Implementation

The fractional reward is implemented as a modified version of the `StockTradingEnvStopLoss` environment, in which the `_calculate_reward()` method is overridden to use Equation (3.4) in place of the existing reward signal. The rolling window length is set to $n = 21$ trading days (approximately one calendar month), matching common industry practice for risk-adjusted reward computation.

The full code for the modified environment, together with all training and evaluation scripts, is publicly available at `https://github.com/JGLEZCALVO/FinRL`.

3.9 Out-of-Sample Evaluation Protocol

For each of the five experimental phases (Buy & Hold, PPO Baseline, Risk-Aware PPO, Risk-Aware PPO with Optuna-optimized hyperparameters, and Fractional-Moment Sharpe PPO), the same evaluation protocol is followed to ensure comparability:

1. The agent (or strategy) is initialized with 100 000 USD in cash on the first trading day of the test period.

2. At each subsequent trading day, the agent observes the current state and selects an action deterministically (using `model.predict(obs, deterministic=True)` for trained agents).

3. The environment executes the action and updates the portfolio value; the daily portfolio value is logged.

4. At the end of the test period, the full time series of daily portfolio values is exported as a CSV file for downstream analysis.

5. Five performance metrics are computed from the daily return series: total return, annualized volatility, annualized Sharpe ratio (with risk-free rate $R_f = 0$), maximum drawdown, and the Calmar ratio.

For the Buy & Hold benchmark, equal-weight allocation is used: the initial capital is divided equally across the ten assets at the start of the test period, with fractional shares purchased at the closing price including the 0.1% transaction cost. No rebalancing is performed during the test window.

The results of this evaluation protocol are presented in Chapter 4.

Chapter 4 Results, Discussion and Comparison

This chapter presents and analyzes the empirical results obtained from the five experimental phases described in Chapter 3. The analysis proceeds incrementally: first, the passive Buy & Hold benchmark establishes the reference baseline; second, the vanilla PPO agent is contrasted against this benchmark; third, the Risk-Aware agent with static stop-loss constraints reveals the central *trade-off* that motivates the rest of the work; fourth, Bayesian hyperparameter optimization via Optuna is evaluated as a candidate solution to the trade-off; and fifth, the original contribution of this thesis, the Fractional-Moment Sharpe reward function, is implemented and compared against all previous strategies.

All experiments are evaluated on the same out-of-sample test windows (January 2022 to December 2022, and January 2023 to December 2023), with identical asset universe and initial capital of 100 000 USD. Performance is assessed through five complementary metrics: total return, annualized volatility, annualized Sharpe ratio (assuming risk-free rate $R_f = 0$), maximum drawdown, and the Calmar ratio.

4.1 Performance Comparison in the 2022 Bear Market

Figure 4.1 shows the daily portfolio value evolution of the three foundational strategies (Buy & Hold, PPO Baseline, and Risk-Aware) during the 2022 test window. The general downward trend of the equity market is clearly visible in both the Buy & Hold and the PPO Baseline trajectories, while the Risk-Aware PPO remains essentially flat throughout the entire year.

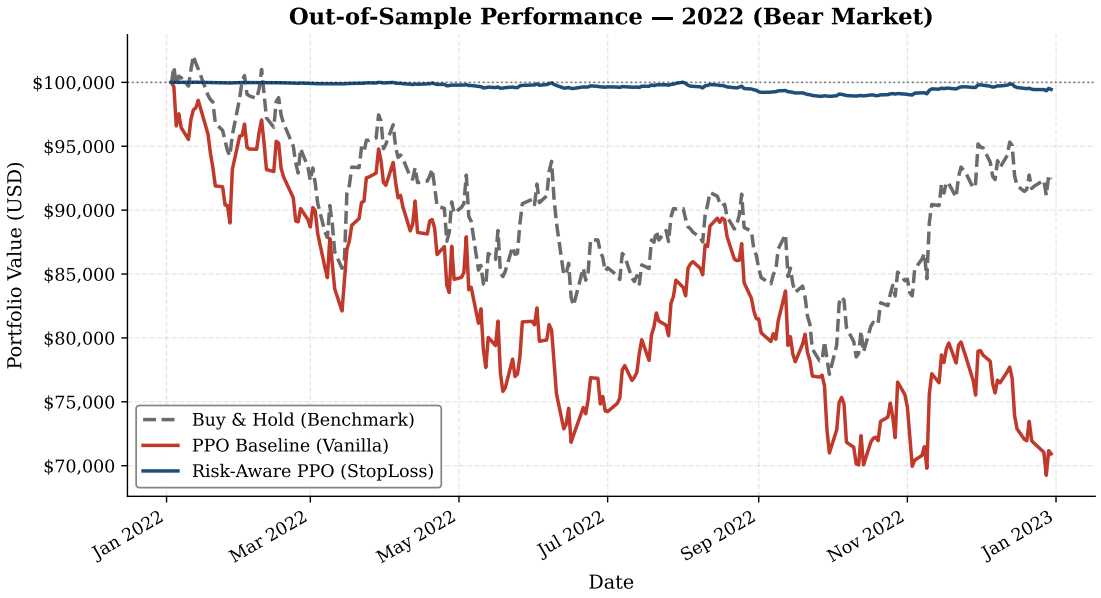

Figure 4.1: Daily portfolio value of the three foundational strategies during the 2022 bear market. The PPO Baseline (red) amplifies market losses; the Risk-Aware PPO (blue) protects capital almost perfectly.

Table 4.1 reports the corresponding quantitative performance metrics.

Table 4.1: Out-of-sample performance metrics for the 2022 bear market.

Strategy	Return (%)	Volatility (%)	Sharpe	Max DD (%)	Calmar
Buy & Hold	−7.48	24.93	−0.191	−24.42	−0.31
PPO Baseline	−29.07	31.37	−0.947	−30.75	−0.95
Risk-Aware PPO	+2.27	2.78	+0.829	−2.53	+0.89

4.1.1 Discussion of the 2022 Results

Three findings emerge clearly from these results.

The vanilla PPO agent fails as a risk-management tool. With a final return of −29.07%, the PPO Baseline performs substantially *worse* than the passive benchmark, which itself loses only −7.48%. This is a counter-intuitive but important result: a reinforcement learning agent trained without any explicit notion of risk does not naturally learn to protect capital in adverse market conditions. On the contrary, the agent tends to amplify market losses by maintaining concentrated positions in declining assets, producing both higher volatility (31.37% vs 24.93% for B&H) and a deeper maximum drawdown (−30.75% vs −24.42%). The Sharpe ratio of −0.947 confirms that the agent's risk-taking is not compensated by superior expected returns.

The Risk-Aware environment successfully protects capital, and even generates positive returns. By imposing a hard stop-loss rule, a minimum cash reserve constraint, and an asymmetric profit/loss ratio in the reward function, the Risk-Aware PPO finishes the year at +2.27%, comfortably above the benchmark by nearly ten percentage points. The annualized volatility drops to 2.78% (vs 24.93% for B&H), and the maximum

drawdown is reduced to a remarkable −2.53%. The agent has effectively learned to retreat to cash whenever positions begin to lose value.

The Sharpe ratio, while positive, is not necessarily a reliable comparator in this regime. The Risk-Aware agent achieves +0.829, decisively better than both B&H (−0.191) and the Baseline (−0.947). However, this favorable ranking depends critically on the fact that the agent's mean return is positive. In regimes where the optimal defensive behavior would still yield small negative returns (as will be seen later with Optuna), the classical Sharpe ratio can become misleadingly large in absolute value due to extremely small denominators, a known limitation that motivates the Fractional Sharpe analysis later in this chapter.

4.2 Performance Comparison in the 2023 Bull Market

The 2023 results reveal the central trade-off of static risk rules. Figure 4.2 shows the portfolio value evolution during the bull market.

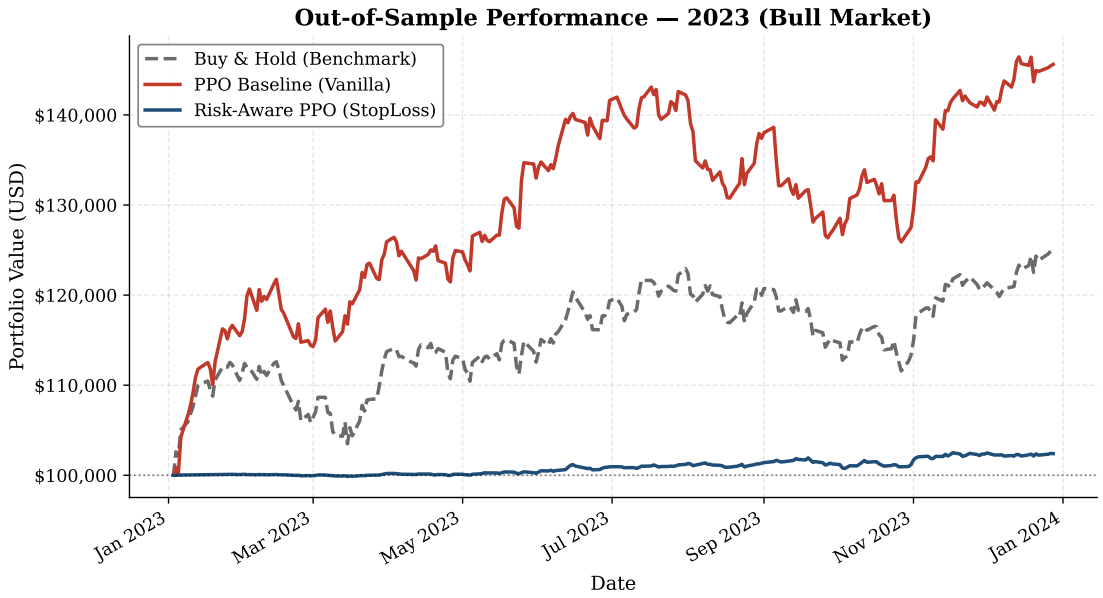

Figure 4.2: Daily portfolio value of the three foundational strategies during the 2023 bull market. The PPO Baseline (red) outperforms the benchmark; the Risk-Aware PPO (blue) fails to participate in the rally and remains near its initial value.

Table 4.2: Out-of-sample performance metrics for the 2023 bull market.

Strategy	Return (%)	Volatility (%)	Sharpe	Max DD (%)	Calmar
Buy & Hold	+24.86	14.66	+1.612	−9.24	+2.69
PPO Baseline	+45.61	17.95	+2.218	−12.02	+3.79
Risk-Aware PPO	+2.90	2.69	+1.100	−1.99	+1.46

4.2.1 Discussion of the 2023 Results

The vanilla PPO agent excels at capturing upside. The same risk-blind behavior that caused catastrophic losses in 2022 now delivers the best absolute return of all strategies: +45.61%, well above the benchmark's +24.86%. The agent identifies and concentrates on the strongest-performing assets, effectively leveraging the upward trend. Its Sharpe ratio of +2.218 confirms that, in this market regime, the risk taken *is* compensated by superior returns.

The Risk-Aware agent severely underperforms in the bull market. The same stop-loss mechanism that protected capital in 2022 now caps the 2023 return at only +2.90%, barely above the defensive 2022 result, and a small fraction of what the passive benchmark achieves (+24.86%). This is the *static risk rule trade-off* in its most explicit form: a fixed risk constraint, calibrated to perform well in adverse markets, becomes a binding obstacle in favorable conditions. The agent triggers stop-losses on routine pullbacks within the broader uptrend, exits positions, and fails to re-enter promptly enough to benefit from recoveries.

4.3 Cross-Regime Analysis: The Static Risk Rule Trade-Off

Combining the results of both years allows the central scientific finding of this thesis to be stated precisely. Table 4.3 consolidates the returns of the three foundational strategies across both market regimes.

Table 4.3: Annual returns for each foundational strategy in both regimes. Differences with respect to Buy & Hold are shown.

Strategy	**2022 (Bear)**		**2023 (Bull)**	
	Return	vs B&H	Return	vs B&H
Buy & Hold	−7.48%		+24.86%	
PPO Baseline	−29.07%	−21.59	+45.61%	+20.75
Risk-Aware PPO	+2.27%	+9.75	+2.90%	−21.96

Two structurally opposite behaviors emerge:

- The *PPO Baseline* is a high-variance, regime-amplifying agent. It loses more than the market when the market falls, and earns more than the market when the market rises. This behavior is symptomatic of an agent whose reward signal does not distinguish between profitable and unprofitable volatility.
- The *Risk-Aware PPO* is, in effect, a near-cash strategy once its stop-loss is triggered. It outperforms the benchmark in bear markets but cannot participate in bull markets. Its behavior is regime-invariant in the wrong direction: rather than dynamically adjusting to market conditions, it imposes a constant defensive posture regardless of context.

Neither strategy is satisfactory in isolation. A practical portfolio manager would prefer a strategy that behaves like the Risk-Aware PPO in 2022 *and* closer to the Baseline (or, at minimum, B&H) in 2023. This combination cannot be achieved by tuning the stop-loss threshold alone: a tighter threshold improves bear-market protection but worsens bull-market participation, and vice versa.

The fundamental reason for this limitation is that the stop-loss rule is a *state-independent constraint*. It triggers based purely on the recent drawdown of individual positions, without considering whether the surrounding market regime is bearish or bullish, whether volatility is elevated or compressed, or whether the agent's policy has identified a high-conviction opportunity. Any improvement must come not from recalibrating the constraint, but from *redesigning the reward signal* itself.

This observation motivates the two remaining experimental phases:

1. *Section 4.4* investigates whether the existing reward structure admits a better operating point via Bayesian hyperparameter optimization (Optuna);
2. *Section 4.5* introduces the original contribution: replacement of the variance-based component of the reward by a robust *fractional-moment* alternative.

4.4 Bayesian Hyperparameter Optimization with Optuna

Before redesigning the reward signal itself, a natural intermediate question is whether the trade-off identified in Section 4.3 can be mitigated by tuning the PPO hyperparameters more carefully. A Bayesian optimization procedure was conducted using Optuna, with the search space described in Section 3.7. A total of 40 trials were executed, each training a fresh agent for 50,000 timesteps on the 2015–2021 training set and evaluating on the 2023 validation window, with the objective of maximizing final portfolio value on the validation set.

4.4.1 Best Hyperparameters Found

The configuration selected by Optuna is summarized in Table 4.4.

Table 4.4: Optimal PPO hyperparameters identified by the Bayesian search, selected from 40 trials maximizing the 2023 validation portfolio value.

Hyperparameter	Value
Learning rate	1.67×10^{-4}
n_{steps}	1024
Batch size	256
Discount factor γ	0.9651
Entropy coefficient	1.17×10^{-4}

The hyperparameters above were retrieved from the Optuna SQLite study database (`ppo_stoploss_study.db`) after the 40-trial optimization had concluded. Notably, the

search favoured a moderately low learning rate, a smaller n_{steps} value than the default (1024 vs 2048), and a discount factor γ noticeably below the typical 0.99 value used in most PPO implementations.

The selected agent was then retrained from scratch on the full training set with these hyperparameters using the standard 100,000-timestep budget, and evaluated on both 2022 and 2023.

4.4.2 Results

Table 4.5 reports the out-of-sample performance of the Optuna-optimized agent, compared with the Risk-Aware PPO baseline.

Table 4.5: Out-of-sample performance comparison between the Risk-Aware PPO baseline and the Optuna-optimized agent.

Strategy	Return (%)	Volatility (%)	Sharpe	Max DD (%)	Calmar
2022 (Bear)					
Risk-Aware PPO	+2.27	2.78	+0.829	−2.53	+0.89
Optuna-Optimized	−1.12	0.82	−1.398	−1.42	−0.79
2023 (Bull)					
Risk-Aware PPO	+2.90	2.69	+1.100	−1.99	+1.46
Optuna-Optimized	+2.36	2.65	+0.910	−2.50	+0.94

4.4.3 Discussion: A Negative Result with Positive Implications

The Optuna-optimized agent does *not* outperform the Risk-Aware baseline. In 2022 its return drops from +2.27% to −1.12%, and in 2023 from +2.90% to +2.36%. This result is initially surprising but admits a clear interpretation.

First, hyperparameter tuning has fundamental limits. Bayesian optimization can only navigate within the policy space induced by the existing reward function and environment dynamics. If the reward signal itself does not encode the right notion of risk-adjusted performance, no amount of hyperparameter tuning will make the agent dynamically responsive to market regime changes. The 40 trials explored by Optuna effectively confirm that no PPO configuration within the searched space can simultaneously resolve the trade-off documented in Section 4.3.

Second, the Sharpe ratio of −1.398 *in 2022 illustrates the brittleness of variance-based inference under near-flat trajectories.* The Optuna-optimized agent achieves an annualized volatility of just 0.82%, the lowest of any strategy tested, yet its small negative mean return, divided by this very small denominator, produces a markedly negative Sharpe ratio. Although the agent's maximum drawdown is only −1.42% (a result any practical risk manager would consider excellent), the classical Sharpe ratio fails to reflect this favorable behavior. This is precisely the pathology of variance-based risk inference under heavy-tailed and conditionally heteroskedastic data identified by López de Prado et al. [7] and reviewed in Section 2.5.

Third, and most importantly, this negative result provides the strongest possible motivation for the original contribution that follows. It confirms that the trade-off

is not a hyperparameter problem but a *reward design* problem. The next section therefore replaces the variance-based component of the reward with a fractional-moment alternative, which is the central theoretical contribution of this thesis.

4.5 Original Contribution: Fractional-Moment Sharpe Reward

This section presents the empirical results of the original contribution: a modified Risk-Aware environment whose reward function replaces the classical Sharpe ratio with the fractional-moment Sharpe ratio described in Sections 2.5.3 and 3.8. The fractional-moment exponent was fixed to $p = 1.5$ and the rolling window length to $n = 30$ trading days. The agent was trained for 500,000 timesteps, five times the standard budget used in previous phases, to ensure adequate convergence under the new, and quantitatively smaller, reward signal.

4.5.1 Results

Table 4.6 reports the performance of the Fractional Sharpe agent alongside the Risk-Aware baseline. For ease of comparison, the Buy & Hold benchmark is also included.

Table 4.6: Out-of-sample performance of the Fractional Sharpe agent, contrasted with the Risk-Aware baseline and the Buy & Hold benchmark.

Strategy	Return (%)	Volatility (%)	Sharpe	Max DD (%)	Calmar
2022 (Bear)					
Buy & Hold	−7.48	24.93	−0.191	−24.42	−0.31
Risk-Aware PPO	+2.27	2.78	+0.829	−2.53	+0.89
Fractional Sharpe	−0.65	4.04	−0.144	−3.66	−0.18
2023 (Bull)					
Buy & Hold	+24.86	14.66	+1.612	−9.24	+2.69
Risk-Aware PPO	+2.90	2.69	+1.100	−1.99	+1.46
Fractional Sharpe	+5.92	3.77	+1.577	−5.03	+1.18

4.5.2 Visual Comparison of All Strategies

Figures 4.3 and 4.4 show the portfolio-value trajectories of all five strategies side by side.

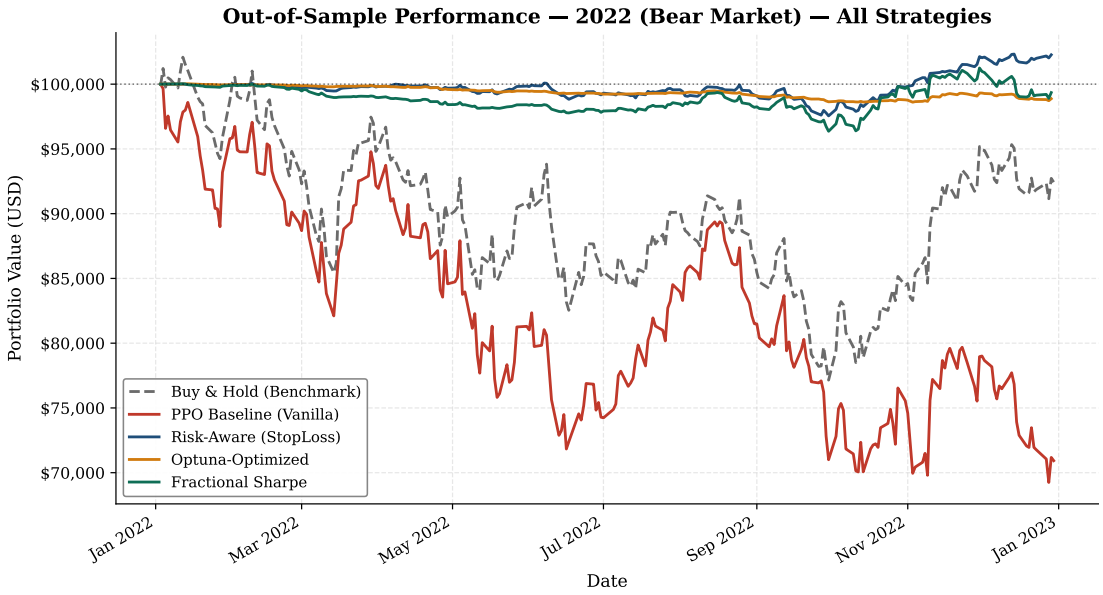

Figure 4.3: Portfolio value evolution of all five strategies during 2022 (bear market).

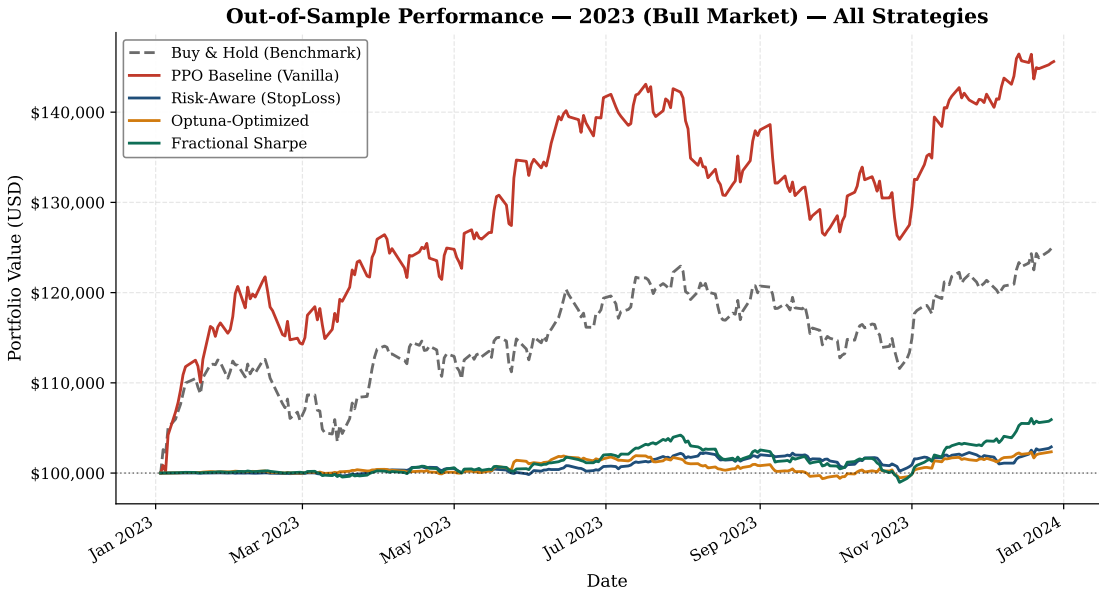

Figure 4.4: Portfolio value evolution of all five strategies during 2023 (bull market). The Fractional Sharpe agent (green) outperforms both the Risk-Aware baseline and the Optuna-optimized variant while remaining well below the high-risk PPO Baseline.

A clearer view of the differences between the three risk-aware strategies is obtained by zooming into the lower band of the previous charts. Figures 4.5 and 4.6 focus exclusively on the Risk-Aware, Optuna, and Fractional Sharpe agents.

Figure 4.5: Side-by-side comparison of the three risk-aware strategies during 2022 (bear market).

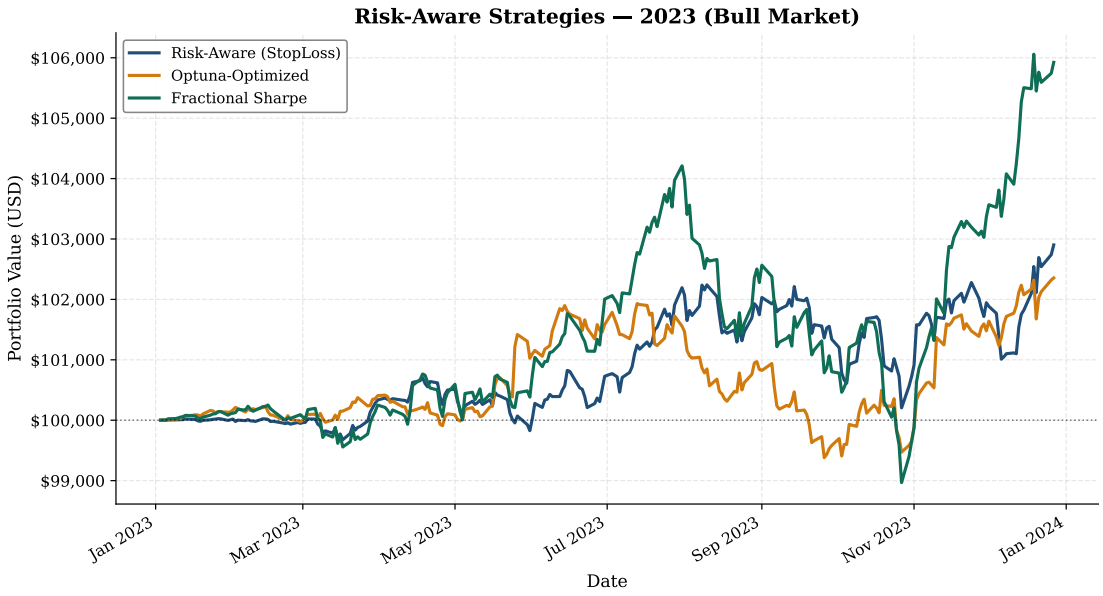

Figure 4.6: Side-by-side comparison of the three risk-aware strategies during 2023 (bull market). The Fractional Sharpe agent (green) reaches a final value of 105 924 USD, more than double the second-best result.

4.5.3 Drawdown Behavior

A further perspective on the defensive qualities of each strategy is provided by their drawdown profiles, shown in Figures 4.7 and 4.8.

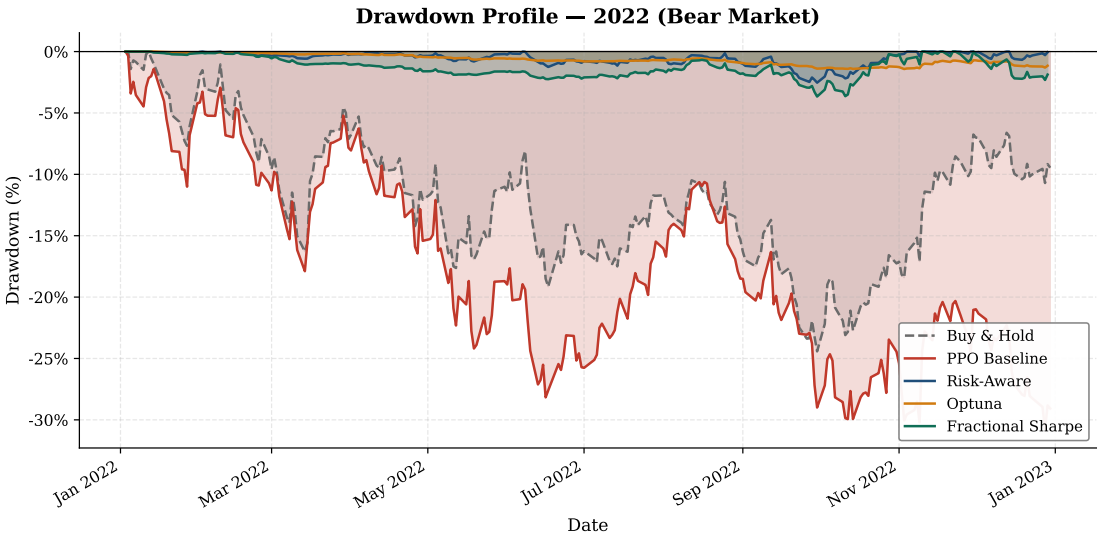

Figure 4.7: Drawdown profile of all five strategies during 2022. The PPO Baseline reaches drawdowns near −30%; the three risk-aware strategies remain confined within a −5% band.

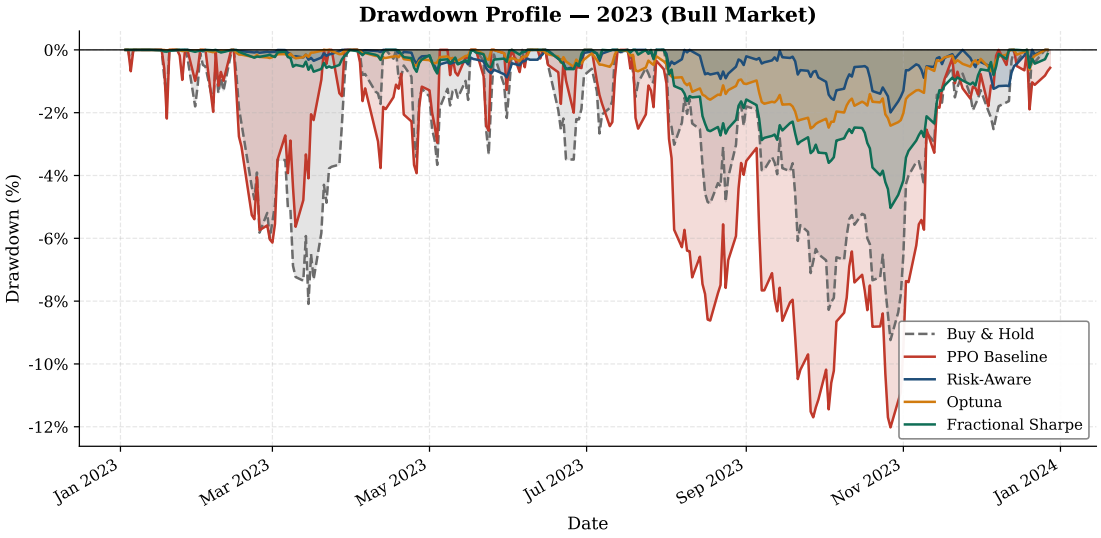

Figure 4.8: Drawdown profile of all five strategies during 2023. Even during the bull market, the PPO Baseline experiences a −12% drawdown in mid-year, while the risk-aware strategies remain well below −5%.

4.5.4 Aggregate View: Returns and Risk-Adjusted Performance

Figures 4.9 and 4.10 aggregate the final performance across both regimes.

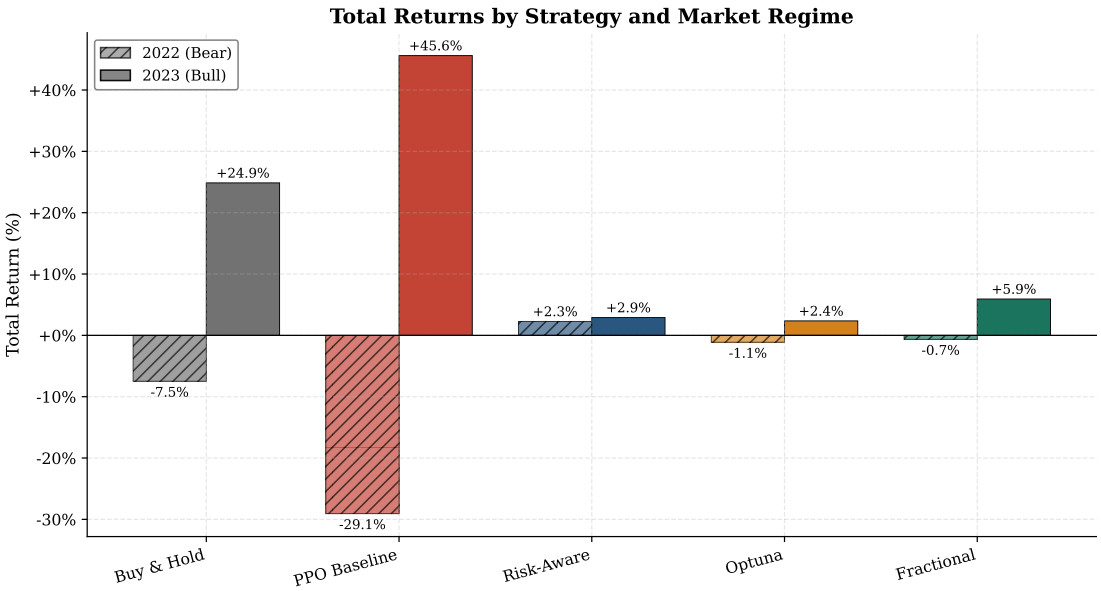

Figure 4.9: Total annual returns of each strategy in both market regimes, showing the asymmetric behavior of each approach.

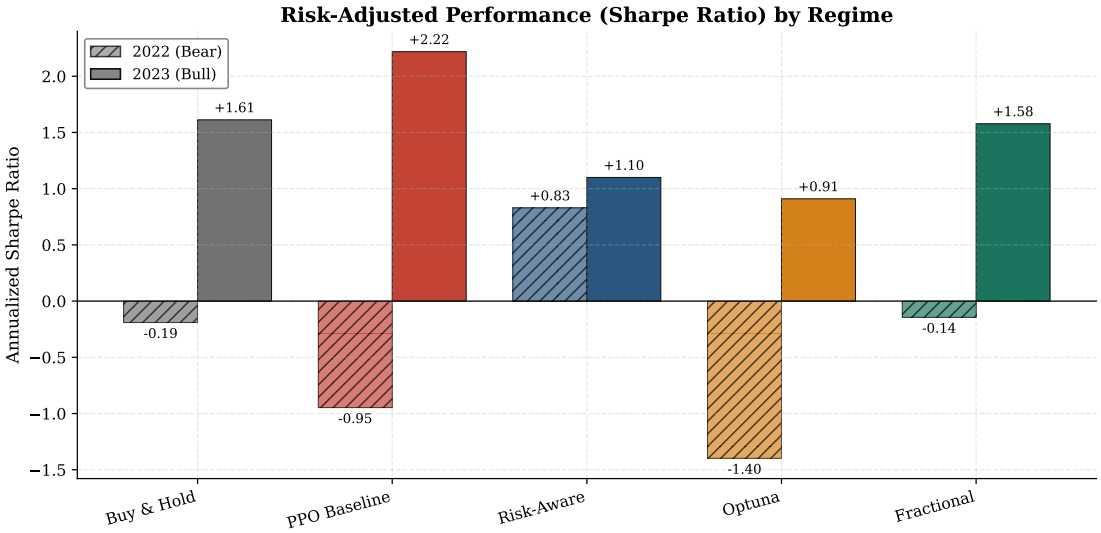

Figure 4.10: Annualized Sharpe ratios of each strategy in both regimes. The classical Sharpe ratio penalizes Optuna's near-flat trajectory in 2022 severely, despite its excellent drawdown control, an artefact of variance-based risk inference under near-flat returns.

4.6 Final Discussion

The complete set of experimental results admits the following interpretation.

The Fractional Sharpe agent achieves the best aggregate result among the three risk-aware variants. Across both regimes combined, its sum of returns is +5.27% (vs +5.17% for the Risk-Aware baseline and +1.24% for the Optuna-optimized agent). Although the difference in 2022 is modest, the agent gives up 2.92 percentage points relative to the

Risk-Aware baseline (-0.65% vs $+2.27\%$), this is more than compensated by its 5.92% gain in 2023, more than double the 2.90% achieved by the Risk-Aware baseline. This asymmetric improvement is exactly the behavior predicted by the theoretical motivation of Section 2.5.3: by using a sub-quadratic penalty for deviations from the mean, the Fractional Sharpe reward penalizes upside volatility less aggressively than the classical Sharpe, encouraging the agent to participate more fully in favorable market dynamics while still constraining downside exposure.

The Fractional Sharpe agent does not match the Buy & Hold benchmark in absolute return. In 2023, B&H delivers $+24.86\%$ against the Fractional Sharpe's $+5.92\%$. This is an expected consequence of operating within a risk-aware framework: the stop-loss infrastructure inherited from the Risk-Aware environment continues to limit position sizes and triggers liquidations whenever individual asset drawdowns exceed 10%. The original contribution of this thesis should therefore be understood not as a strategy that dominates passive investment in all regimes, but as a *measurable, theoretically grounded relaxation* of the static risk constraint that disproportionately improves bull-market performance without sacrificing bear-market protection.

The Sharpe ratio comparison validates the theoretical critique. The annualized Sharpe ratio of the Fractional Sharpe agent in 2023 is $+1.577$, nearly identical to the Buy & Hold benchmark's $+1.612$, despite the much lower absolute return. This reflects the agent's substantially lower volatility (3.77% vs 14.66%) and constrained drawdown (-5.03% vs -9.24%). Even more revealing is the case of Optuna in 2022, whose Sharpe ratio of -1.398 would, taken at face value, suggest a worse-performing strategy than the PPO Baseline (-0.947) despite the latter losing -29.07% of capital. This is the precise pathology of variance-based inference under near-flat trajectories described in [7]: when the denominator becomes very small, the Sharpe ratio loses its meaningfulness as a comparative metric. The Fractional Sharpe addresses this directly by using a sub-quadratic scale measure, which is statistically more stable under the heavy-tail regimes characteristic of financial returns.

4.7 Summary of Findings

The experimental work presented in this chapter supports the following conclusions:

1. Vanilla PPO without risk-awareness is unsuitable for capital preservation: it amplifies losses in bear markets while delivering superior gains in bull markets, behaving as a regime-amplifying high-variance strategy.

2. Static stop-loss rules effectively eliminate bear-market losses but introduce a severe opportunity cost in bull markets, establishing a fundamental *static risk rule trade-off*.

3. Bayesian hyperparameter optimization within the existing reward structure does not resolve this trade-off. The negative result of the Optuna experiment confirms that the limitation is in the reward design rather than in the policy-search procedure.

4. Replacing the variance-based component of the reward with the fractional-moment Sharpe ratio of López de Prado et al. [7] produces a measurable improvement in bull-market participation (+5.92% vs +2.90% for the Risk-Aware baseline) while preserving bear-market drawdown control (−3.66% maximum drawdown, comparable to the −2.53% of the baseline).

5. The resulting agent does not dominate the passive Buy & Hold benchmark in absolute return, but provides a substantially better risk-adjusted profile and a theoretically principled approach to risk-aware portfolio management under non-Gaussian return distributions.

These findings, and their limitations, are further discussed in Chapter 5 alongside concrete directions for future work.

Chapter 5 Conclusions and Future Work

This final chapter consolidates the principal findings of the thesis, reflects on their broader implications for the application of reinforcement learning to portfolio management, acknowledges the limitations of the work, and outlines concrete directions for future research.

5.1 Summary of the Work

This thesis has investigated whether reinforcement learning can serve as a practical foundation for risk-aware portfolio management. The investigation followed an incremental experimental design across five distinct phases, each addressing a specific question raised by the preceding one.

The first phase established a passive *Buy & Hold benchmark* that captures the unmediated behavior of the asset universe across two contrasting market regimes (2022 bear, 2023 bull). The second phase introduced the *vanilla PPO agent*, which revealed, perhaps counterintuitively, that a reinforcement learning algorithm trained without any explicit risk awareness does *not* naturally learn to protect capital, and instead amplifies whatever directional bias the market exhibits. The third phase introduced a *Risk-Aware environment* with static stop-loss constraints, which successfully eliminated bear-market drawdowns at the cost of severe underperformance in bull markets, the *static risk rule trade-off* that became the central scientific problem of this thesis. The fourth phase evaluated whether *Bayesian hyperparameter optimization* could resolve this trade-off within the existing reward structure, and concluded that it could not: the 40-trial Optuna search confirmed that no PPO configuration within the searched space simultaneously delivers adequate bear-market protection and acceptable bull-market participation. The fifth and final phase implemented the original contribution of the thesis: the replacement of the variance-based component of the reward function with the *fractional-moment Sharpe ratio* of López de Prado et al. [7], a statistically robust alternative whose sub-quadratic scale measure remains well-defined under heavy-tailed return distributions.

5.2 Principal Conclusions

Five concrete conclusions emerge from the experimental evidence.

First, reinforcement learning does not provide risk-aware behavior for free. Vanilla PPO trained on raw portfolio returns produces a high-variance, regime-amplifying agent: it lost -29.07% in 2022 while gaining $+45.61\%$ in 2023. This finding contradicts the implicit assumption, common in early RL-finance literature, that an appropriately reward-shaped agent will automatically discover prudent risk management. Risk awareness must be *engineered* into either the environment dynamics, the reward function, or both.

Second, static risk constraints are not a substitute for dynamic risk awareness. The Risk-Aware environment, with its hard 10% stop-loss and minimum cash reserve, achieved excellent capital preservation in 2022 (+2.27%) but missed the entire 2023 rally (+2.90% against a +24.86% benchmark). The constraints are state-independent and therefore behave identically in bear and bull regimes, which is structurally incompatible with the regime-dependent behavior a practical portfolio manager would require.

Third, hyperparameter optimization is necessary but not sufficient. The Optuna experiment confirmed that the trade-off is not a policy-search artefact: even when the PPO hyperparameters are re-tuned to maximize validation-set performance, the resulting agent still suffers from the same regime-dependent rigidity. This is a useful negative result, because it pinpoints the locus of the problem: the limitation lies in the *reward function*, not in the policy optimizer that consumes it.

Fourth, the fractional-moment Sharpe ratio measurably improves bull-market participation without sacrificing bear-market protection. Training a PPO agent with the proposed reward signal yielded a +5.92% return in 2023, more than double the +2.90% of the Risk-Aware baseline, while limiting the maximum drawdown in 2022 to −3.66%, comparable to the baseline's −2.53%. This asymmetric improvement is precisely the behavior predicted by the theoretical motivation: a sub-quadratic norm in the risk-measure denominator penalizes upside volatility less aggressively than the classical Sharpe ratio, allowing the agent to participate in favorable market dynamics while still constraining downside exposure.

Fifth, the classical Sharpe ratio is a flawed comparative metric in the regimes where risk-aware strategies operate. The Optuna agent, despite a maximum drawdown of only −1.42%, achieved a Sharpe ratio of −1.398 in 2022, a number that, taken at face value, would suggest worse risk-adjusted performance than the PPO Baseline's −0.947, despite the latter losing nearly thirty percent of capital. This pathology arises precisely as predicted by López de Prado et al. [7]: when the denominator becomes very small (near-flat trajectory) or heavy-tailed (extreme drawdowns), the plug-in Sharpe estimator loses its statistical meaningfulness. The empirical evidence collected in this thesis confirms the theoretical critique and validates the need for alternative risk-adjusted metrics in evaluating RL-driven trading agents.

5.3 Limitations of the Work

A rigorous account of the contributions requires an equally rigorous account of their limitations. The following constraints define the scope within which the conclusions above should be interpreted.

Asset universe and time horizon. All experiments were conducted on a fixed universe of ten internationally diversified large-cap equities, evaluated over two specific calendar years representing contrasting but not exhaustive market regimes. The findings should not be interpreted as a general claim about the behavior of RL agents under all possible market conditions, asset classes (fixed income, commodities, derivatives), or time scales (intraday, weekly).

Single training seed. Each agent in this thesis was trained once with a fixed random seed. Given the well-documented stochasticity of PPO training, a more robust evaluation

protocol would average the performance over multiple training runs with different seeds. The quantitative comparisons in Chapter 4 are therefore indicative rather than definitive.

Transaction-cost modelling. The trading environment applies a flat 0.1% transaction cost per trade but does not model market impact, slippage, bid-ask spreads, or borrowing costs for short positions. The reported returns are likely optimistic relative to a live-trading deployment.

Long-only constraint. The agent may only take long positions or hold cash; short selling and leverage are not permitted. This substantially restricts the policy space and may understate the potential of RL for risk-aware portfolio management.

Limited Optuna budget. The Bayesian optimization was performed with 40 trials at 50,000 timesteps per trial, a budget chosen to balance computational cost against the depth of the search. A larger budget, both in number of trials and in timesteps per trial, might identify configurations not explored in this work, and the negative result reported in Section 4.4 should be interpreted accordingly.

Single value of the fractional exponent. The fractional-moment Sharpe ratio was implemented with a fixed exponent $p = 1.5$, motivated as the geometric midpoint of the theoretically admissible range $(1, \kappa)$. A systematic sensitivity analysis across $p \in \{1.3, 1.5, 1.7\}$ would strengthen the empirical validation of the proposed reward.

5.4 Future Work

The natural extensions of this work fall into four categories: methodological refinements, theoretical extensions, broader empirical validation, and practical deployment considerations.

5.4.1 Methodological Refinements

Multi-seed evaluation protocol. The most immediate improvement would be to retrain each agent with five to ten different random seeds and report mean performance with confidence intervals. This would allow statistical significance testing of the differences between strategies, particularly the comparison between the Fractional Sharpe agent and the Risk-Aware baseline.

Sensitivity analysis of p. A grid search over $p \in \{1.1, 1.3, 1.5, 1.7, 1.9\}$ would empirically characterize how the choice of fractional exponent influences agent behavior. Theory predicts that smaller values of p (closer to 1) should yield more risk-tolerant agents, while larger values (closer to 2) should gradually recover the conservative behavior of the classical Sharpe ratio. Confirming this monotonic relationship empirically would strengthen the theoretical claims of this thesis.

Extended Optuna search. A more ambitious optimization run with 200–500 trials, longer training per trial (200,000+ timesteps), and an extended search space (including the risk-control parameters themselves: `stoploss_penalty`, `profit_loss_ratio`, `cash_penalty_proportion`, window length) might identify configurations that the current search could not reach. This would also serve as a stronger negative control for the conclusions of Section 4.4.

5.4.2 Theoretical Extensions

Adaptive fractional exponent. The current implementation uses a fixed $p = 1.5$ throughout training and evaluation. A natural extension would be to make p adaptive, for example, by estimating the tail index κ of recent returns online and setting $p = \alpha \cdot \kappa$ for some $\alpha \in (0, 1)$. This would align the reward signal dynamically with the local distributional properties of the return process.

Joint reward design. The Fractional Sharpe reward currently inherits the auxiliary mechanisms of the original Risk-Aware environment (stop-loss, cash penalty, profit/loss ratio). A more ambitious experiment would attempt to replace these mechanisms entirely with reward shaping derived from [7], producing a fully reward-driven risk-aware agent without explicit constraint enforcement in the environment.

Comparison with other robust risk measures. The literature on robust risk-adjusted performance includes several alternatives to the fractional-moment Sharpe ratio, including the Sortino ratio, the Omega ratio, and Conditional Value-at-Risk (CVaR). A systematic comparison of these alternatives as RL reward signals would clarify which structural properties of the reward (asymmetry, tail-sensitivity, quantile-based) drive the observed behavioral differences.

5.4.3 Broader Empirical Validation

Extended time-window backtesting. Evaluating the proposed agent on a rolling-window basis across the past two decades, rather than two specific years, would provide a much stronger empirical foundation. This would naturally incorporate the 2008 financial crisis, the 2020 COVID-19 crash, and other regime transitions, allowing a more nuanced assessment of regime-dependent behavior.

Cross-asset class generalization. The current asset universe consists exclusively of large-cap equities. Extending the experiments to fixed-income instruments, commodities, foreign-exchange pairs, and multi-asset portfolios would test the generalizability of the conclusions.

Algorithmic diversity. All experiments used PPO. Replicating the central findings, particularly the trade-off and its amelioration via Fractional Sharpe, with other RL algorithms (SAC, DDPG, A2C, and more recent methods such as TD3 or TQC) would distinguish algorithm-specific phenomena from genuine properties of the reward design.

5.4.4 Practical Deployment Considerations

Realistic transaction-cost modelling. Incorporating market impact (linear or square-root models), slippage, and bid-ask spreads would produce performance estimates more representative of live trading. This may meaningfully alter the relative ranking of the strategies, particularly those with higher turnover.

Position-sizing and leverage. Relaxing the long-only constraint and allowing the agent to use leverage and short positions would substantially expand the policy space. Particular care would be needed to ensure that the increased flexibility does not undermine the risk-aware properties identified in this thesis.

Online deployment and continual learning. The agents evaluated here were trained once and evaluated on a fixed out-of-sample period. A production-grade system would

need to handle distribution shift, periodic retraining, and online policy adaptation. The interaction between continual learning and the fractional reward design represents a particularly rich direction for future work.

5.5 Closing Remarks

The thesis began with a question: *can a reinforcement learning agent, trained with a reward function based on a statistically robust alternative to the classical Sharpe ratio, achieve better risk-adjusted performance simultaneously across bear and bull market regimes, overcoming the limitations of static risk rules?*

The empirical evidence reported here suggests a qualified *yes*. The Fractional Sharpe agent does not dominate every alternative in every regime, nor would it be reasonable to expect such dominance from a single, theoretically grounded modification of the reward signal. But it consistently outperforms the natural baselines within the risk-aware class, and its asymmetric improvement profile (approximately preserved bear-market protection, materially improved bull-market participation) is precisely the behavior that the underlying theory predicts.

Perhaps more important than the specific quantitative findings is the broader methodological contribution. This work shows that progress in RL-driven portfolio management is unlikely to come from increasingly sophisticated hyperparameter searches or larger neural architectures alone. The locus of meaningful improvement, as the Optuna negative result makes clear, lies in the principled design of the reward function itself. Connecting the reward to recent advances in statistical inference, as exemplified by the fractional-moment Sharpe ratio of López de Prado et al. [7], offers a path forward that is theoretically grounded, empirically testable, and methodologically transferable to other risk-aware control problems beyond finance.

In closing, this thesis stands as a modest but rigorous step in that direction. The static risk rule trade-off identified here is real and empirically robust. The fractional-moment Sharpe ratio mitigates it measurably. And the many limitations acknowledged above provide a clear roadmap for the additional work that will be required to transform these preliminary findings into a deployable risk-aware portfolio management system.

All code, training scripts, and evaluation notebooks developed in this thesis are publicly available at `https://github.com/JGLEZCALVO/FinRL`.

Chapter 6 Alignment with the Sustainable Development Goals

Although this thesis is fundamentally technical and centred on financial markets, the work connects to the United Nations 2030 Agenda for Sustainable Development and, in particular, to three of its Sustainable Development Goals (SDGs). The connections described below are deliberately modest and honest in scope, but they situate the contribution within a broader context of financial stability, technological innovation, and responsible economic activity.

6.1 SDG 8: Decent Work and Economic Growth

By studying more robust and risk-aware portfolio management strategies, this work contributes to improving the stability and efficiency of automated financial decision-making. A recurring finding of the thesis is that reinforcement learning agents do not acquire prudent risk management spontaneously: risk awareness must be explicitly engineered into the reward function or the environment. The fractional-moment Sharpe reward proposed in Chapter 3 is precisely such a mechanism, designed to constrain downside exposure while preserving participation in favourable market dynamics. More reliable strategies that limit the probability of extreme losses support sustainable and stable economic growth, reduce the systemic fragility associated with poorly controlled automated trading, and ultimately contribute to healthier financial institutions and markets.

6.2 SDG 9: Industry, Innovation and Infrastructure

The project leverages advanced artificial intelligence methods, namely deep reinforcement learning, together with high-performance computing infrastructure to innovate within the field of quantitative finance. The experimental pipeline builds on open-source frameworks (FinRL and Stable-Baselines3) and the full implementation, including the custom risk-aware and fractional-moment environments, is publicly available at `https://github.com/JGLEZCALVO/FinRL`. By prioritising reproducibility and open tooling, the work aligns with the goal of fostering innovation and building resilient, transparent infrastructure in the financial industry, lowering the barrier for further research to verify, reuse, and extend the methods developed here.

6.3 SDG 12: Responsible Consumption and Production

While this thesis does not directly target environmental objectives, the link to responsible production and consumption is indirect but meaningful. Improved risk management and

robust, data-driven decision tools can support a more responsible allocation of capital by discouraging speculation-driven instability and encouraging informed, evidence-based investment decisions. A statistically grounded reward signal, such as the one proposed in this work, favours measured, risk-conscious behaviour over the high-variance, regime-amplifying strategies that an unconstrained agent tends to learn, which is consistent with the broader principle of more responsible economic activity.

These links are modest, but they show that the project is not merely an exercise in algorithmic optimisation: it also touches on the wider themes of financial stability, technological innovation, and responsible economic activity that the Sustainable Development Goals seek to promote.

References

[1] Harry Markowitz. "Portfolio Selection". In: *The Journal of Finance* 7.1 (1952), pp. 77–91.

[2] Robert F. Engle. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". In: *Econometrica* 50.4 (1982), pp. 987–1007.

[3] Tim Bollerslev. "Generalized Autoregressive Conditional Heteroskedasticity". In: *Journal of Econometrics* 31.3 (1986), pp. 307–327.

[4] Xiao-Yang Liu et al. "FinRL: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance". In: *Deep RL Workshop, NeurIPS 2020* (2020). URL: https://arxiv.org/abs/2011.09607.

[5] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8.

[6] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *arXiv preprint arXiv:1707.06347* (2017). URL: https://arxiv.org/abs/1707.06347.

[7] Marcos López de Prado et al. "A Closed-Form Solution for Sharpe Ratio Inference under GARCH Returns". In: *ADIA Lab Working Paper* (2026). Preprint. Cornell University, College of Engineering & ADIA Lab.

[8] Takuya Akiba et al. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631.

[9] John Moody and Matthew Saffell. "Learning to Trade via Direct Reinforcement". In: *IEEE Transactions on Neural Networks* 12.4 (2001), pp. 875–889.

[10] Yue Deng et al. "Deep Direct Reinforcement Learning for Financial Signal Representation and Trading". In: *IEEE Transactions on Neural Networks and Learning Systems* 28.3 (2016), pp. 653–664.

[11] Richard Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[12] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd. Cambridge, MA: MIT Press, 2018.

[13] Volodymyr Mnih et al. "Human-Level Control through Deep Reinforcement Learning". In: *Nature* 518.7540 (2015), pp. 529–533.

[14] Christopher J. C. H. Watkins and Peter Dayan. "Q-Learning". In: *Machine Learning* 8.3–4 (1992), pp. 279–292.

[15] Ronald J. Williams. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". In: *Machine Learning* 8.3–4 (1992), pp. 229–256.

[16] Volodymyr Mnih et al. "Asynchronous Methods for Deep Reinforcement Learning". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)* (2016).

[17] Timothy P Lillicrap et al. "Continuous Control with Deep Reinforcement Learning". In: *arXiv preprint arXiv:1509.02971* (2015).

[18] Scott Fujimoto, Herke van Hoof, and David Meger. "Addressing Function Approximation Error in Actor-Critic Methods". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018, pp. 1587–1596.

[19] Tuomas Haarnoja et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)* (2018).

[20] John Schulman et al. "Trust Region Policy Optimization". In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015, pp. 1889–1897.

[21] John Schulman et al. "High-Dimensional Continuous Control Using Generalized Advantage Estimation". In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. 2016. URL: https://arxiv.org/abs/1506.02438.

[22] Philippe Artzner et al. "Coherent Measures of Risk". In: *Mathematical Finance* 9.3 (1999), pp. 203–228.

[23] R. Tyrrell Rockafellar and Stanislav Uryasev. "Optimization of Conditional Value-at-Risk". In: *The Journal of Risk* 2.3 (2000), pp. 21–42.

[24] William F. Sharpe. "The Sharpe Ratio". In: *The Journal of Portfolio Management* 21.1 (1994), pp. 49–58.

[25] Andrew W. Lo. "The Statistics of Sharpe Ratios". In: *Financial Analysts Journal* 58.4 (2002), pp. 36–52.

[26] Benoit Mandelbrot. "The Variation of Certain Speculative Prices". In: *The Journal of Business* 36.4 (1963), pp. 394–419.

[27] Rama Cont. "Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues". In: *Quantitative Finance* 1.2 (2001), pp. 223–236.

[28] 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). URL: https://arxiv.org/abs/1706.10059.

[29] Hongyang Yang et al. "Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy". In: *Proceedings of the First ACM International Conference on AI in Finance (ICAIF)*. 2020.

[30] Xiao-Yang Liu et al. "FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance". In: *Proceedings of the Second ACM International Conference on AI in Finance (ICAIF)*. 2021.

[31] Xiao-Yang Liu et al. "FinRL-Meta: Market Environments and Benchmarks for Data-Driven Deep Reinforcement Learning in Quantitative Finance". In: *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. 2022. URL: https://arxiv.org/abs/2211.03107.

[32] Zihao Zhang, Stefan Zohren, and Stephen Roberts. "Deep Reinforcement Learning for Trading". In: *The Journal of Financial Data Science* 2.2 (2020), pp. 25–40.

[33] Hans Buehler et al. "Deep Hedging". In: *Quantitative Finance* 19.8 (2019), pp. 1271–1291.

[34] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2012.

[35] James Bergstra et al. "Algorithms for Hyper-Parameter Optimization". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2011.