REINFORCEMENT LEARNING FOR ALGORITHMIC TRADING

What are the impacts of different possible reward functions on the ability of the RL model to learn, and the performance of the RL Model?

Author: Justas Bertasius
Supervisors: Antonis Papapantoleon, Amin Sharifi Kolarijani
Responsible Professor: Neil Yorke-Smith

TU Delft
Delft University of Technology

1) BACKGROUND

Algorithmic trading [1] replaces human discretion with rule-driven code that can scan live market feeds and shoot orders in microseconds. These **systems thrive on speed, scalability, and emotion-free consistency.**

Reinforcement learning (RL) trains an agent to interact with an environment, choosing actions that maximise long-run reward via trial and error. The formalism is a **Markov Decision Process (MDP)** $M = (S, A, P, R, \gamma)$ [2]. Key RL elements include:

- **Agent** – decision-maker.
- **Environment** – market simulator.
- **State** – the position the agent is currently in.
- **Action** – trade (e.g. Buy, Sell, Hold).
- **Reward** – profit, risk metric, etc.

2) RESEARCH QUESTION

What are the impacts of different possible reward functions on the ability of the RL model to learn, and the performance of the RL Model?

Subquestions:

- **SQ1: Profit-only vs. risk-adjusted rewards**. Does replacing **PnL based** rewards with **risk-adjusted** ones yield better performance?
- **SQ2: Multi-objective vs single-objective rewards:** Does using **multi-objective** rewards (weighted combination of **profit**, **risk**, **transaction costs** and **drawdown penalty**) improve the performance of the model and how do these components of the multi-objective reward function impact the performance?
- **SQ3: Self-rewarding mechanism**: Does a **self-rewarding** [3] mechanism improve adaptability of the model, thus resulting in better results?
- **SQ4: Imitation learning:** Does **imitation learning** [4] helps to compute a reward function which would improve the performance of the model?

3) METHODOLOGY

- **Data.** 15 minute EUR/USD currency pair data was downloaded from the **Dukascopy** public archive [5] covering the period 2 Jan. 2022 – 16 May 2025. It was then split into **train** and **evaluation** sets with ratios 0.7 and 0.3 (Figure 1 and 2)

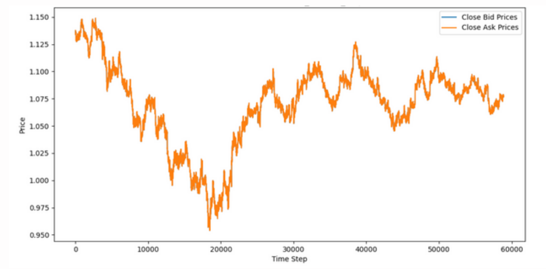


Figure 1: Market training data

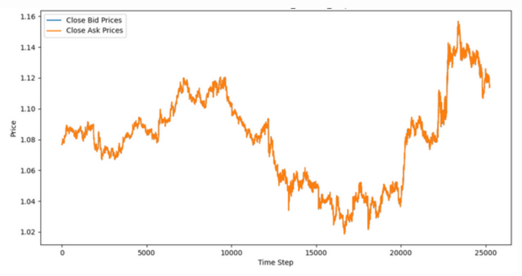
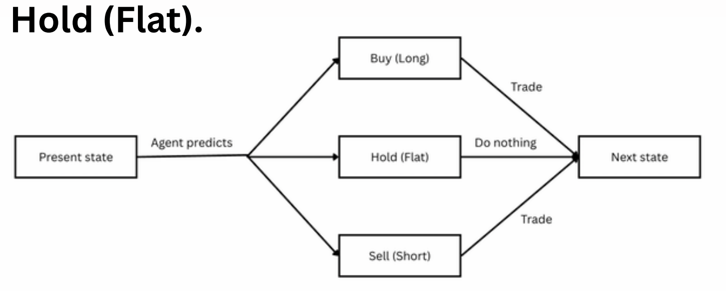


Figure 2: Market evaluation. data

- **Trading environment.** A custom newly created gymnasium environment, **ForexEnv**, emulates leveraged spot trading under realistic frictions while retaining analytical clarity. It consists of three actions: **Buy (Long), Sell (Short)** and **Hold (Flat).**



- **Agent.** A **Deep Q-Network (DQN)** agent consisting of **two hidden layers** of 128 neurons with **ReLU** activations, **learning rate** 10^-4, **replay buffer** 50 000, **batch** 64, **target net sync** every 500 steps, γ =0.99.
- **Experiment procedure.** The explored reward functions are evaluated under **identical fixed training settings,** using 50 **episodes** per run over 5 different **seeds** and **the best performing** model over all the episodes is taken and analyzed further.
- **Evaluation metrics.** Cumulative return, Sharpe ratio, maximum drawdown, profit factor, trade win rate, reward evolution and action distribution.

4) RESULTS AND DISCUSSION

Figure 3: Profit-based rewards total returns

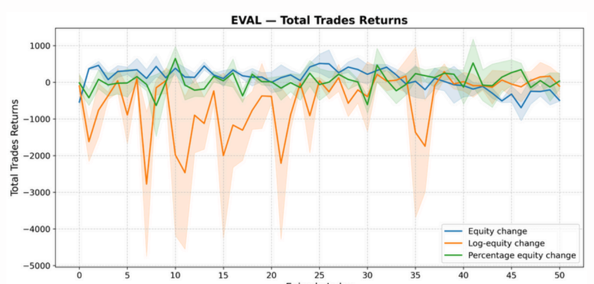


Figure 4: Risk-adjusted rewards total returns

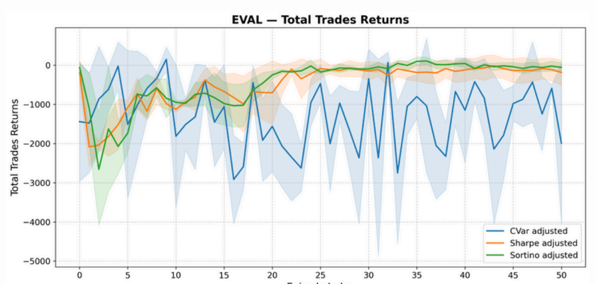


Figure 5: Multi-objective rewards total returns

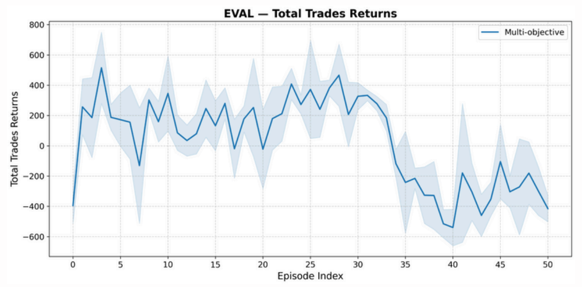


Figure 6: Self-rewarding based rewards total returns

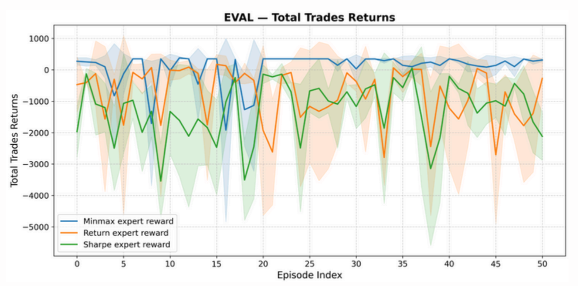
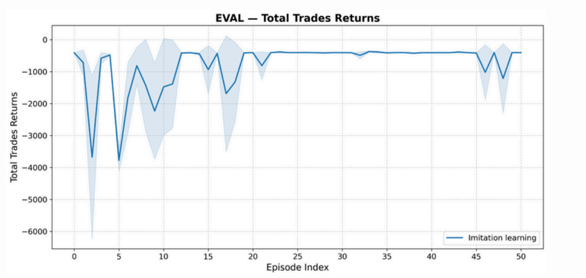


Figure 7: Imitation learning based rewards total returns



- **Profit-based.** Figure 3 shows that the equity-change agent learns most smoothly, whereas the log- and percentage variants display substantial early-stage volatility. After episode 30, the equity-change curve plateaus, indicating possible over-fitting.
- **Risk-adjusted.** Figure 4 reveals that the CVaR agent is highly erratic, whereas the Sharpe- and Sortino-adjusted agents learn more smoothly and exhibit much smaller deviation.
- **Multi-objective.** As shown in Figure 5, training is profitable and relatively stable until roughly episode 30, at which point returns deteriorate, suggesting sensitivity to the fixed weight vector or overfitting.
- **Self-rewarding.** Figure 6 shows that the min-max agent converges most smoothly; the other two variants remain volatile throughout.
- **Imitation learning.** Figure 7 shows large fluctuations during the first 20 episodes, followed by convergence to consistently small, often negative, returns.

Figure 8: Equity curve for profit-based rewards

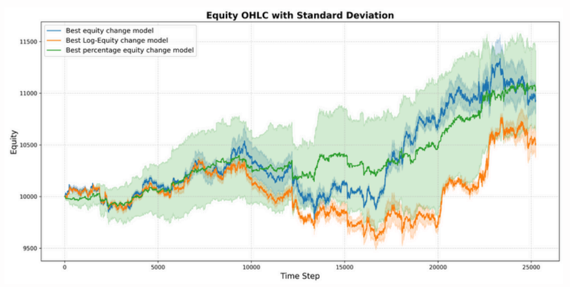


Figure 9: Equity curve for risk-adjusted rewards

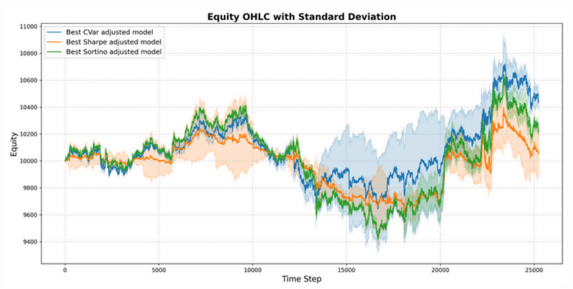


Figure 10: Equity curve for multi-objective rewards

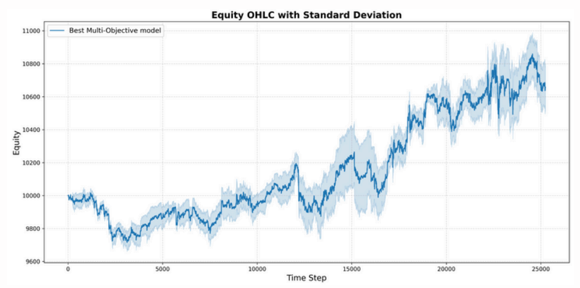


Figure 11: Equity curve for self-rewarding based rewards

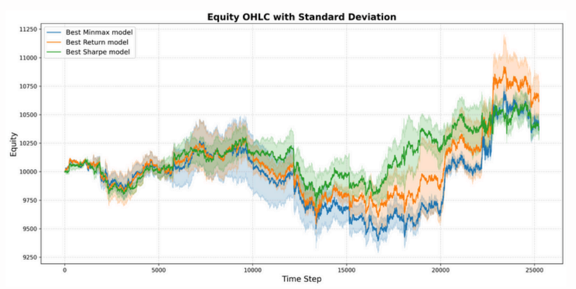


Figure 12: Equity curve for imitation learning based rewards

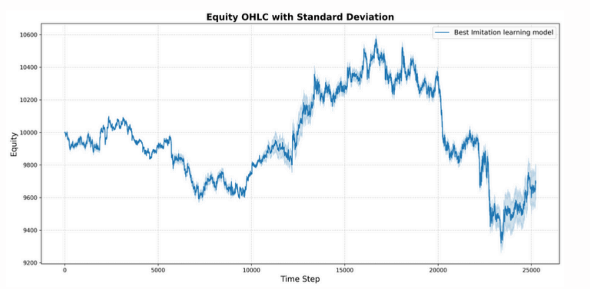


Table 1: Performance metrics of the best model for every reward function variation

Metric	Equity-change		Sortino-adj.		Multi-obj.		Self-reward (Return)		Imitation	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Sharpe Ratio ↑	1.141	0.174	0.303	0.042	0.956	0.279	0.774	0.227	−0.340	0.118
Max Drawdown ↑	−764.86	108.37	−1000.53	15.04	−395.83	38.69	−862.07	143.68	−1150.61	268.44
Profit Factor ↑	1.024	0.004	1.006	0.001	1.025	0.011	1.016	0.005	0.994	0.002
Equity Change (%) ↑	9.27	1.47	2.17	0.38	6.58	1.11	6.26	2.00	−3.07	1.19
Total Trades Returns ↑	926.89	146.97	217.03	37.59	658.31	111.08	626.12	200.04	−306.72	122.89
Total Trades	186	76	273	46	959	123	353	589	53	65
Long Trades Count	92	32	224	98	501	86	177	294	5	5
Short Trades Count	94	44	49	56	458	50	176	293	48	60
Win Rate (%) ↑	66.17	3.69	37.62	2.83	65.26	0.88	70.24	3.83	50.33	0.49
Avg Rewards ↑	0.037	0.006	0.178	0.003	0.0145	0.0087	0.458	0.074	-2.9×10^{-6}	3.6×10^{-6}

- **Profit-based.** Figure 8 shows equity change reward model being the most stable with slightly smaller return than percentage variant.
- **Risk-adjusted.** Figure 9 shows that the CVaR model occasionally achieves larger returns, however its instability makes the Sortino-adjusted model the preferred choice.
- **Multi-objective.** Figure 10 depicts the equity path for the best multi-objective model. Although its cumulative profit is below that of the pure profit agent, drawdowns are substantially lower.
- **Self-rewarding.** Figure 11 displays equity curves for the best model of each variant. Despite its instability, the return-expert agent delivers the highest total return.
- **Imitation learning.** Figure 12 shows that even though imitation learning model is table it does not perform well and almost moves inversely with the market..

5) CONCLUSIONS

- **SQ1: Profit-based** rewards outperformed **risk-adjusted** ones in almost every metric.
- **SQ2: Multi-objective** rewards performed decently, but still worse than **profit-based** ones.
- **SQ3: Self-rewarding based** rewards performed quite poorly possibly due to being too complex and needing more data and in depth fine-tuning
- **SQ4: Imitation learning based** rewards performed by far the worst possibly due to also being too complex.

6) FUTURE WORK

- **Broader environments:** Add multiple assets, regime shifts, slippage, and fees.
- **Reward–algorithm fit:** Test how each reward pairs with alternative agents, features, and exploration schemes.
- **Hyper-parameter tuning:** Search learning rates, buffer sizes, and reward weights for faster convergence.
- **Advanced architectures:** Policy-gradient and actor–critic models might help to produce better results on some rewards.
- **Adaptive rewards:** Let multi-objective weights adjust on-line to market conditions.
- **Live evaluation:** Deploy agents in paper-trading or realistic simulators to test out-of-sample robustness.

7) REFERENCES

[1] Sachin Napate, Mukul Thakur, and D B-Schlool. Algorithmic trading and strategies. November 2020.
[2] Martijn Otterlo and Marco Wiering. Reinforcement learning and markov decision processes. Reinforcement Learning: State of the Art, pages 3–42, 01 2012.
[3] Yuling Huang, Chujin Zhou, Lin Zhang, and Xiaoping Lu. A self-rewarding mechanism in deep reinforcement learning for trading strategy optimization. Mathematics, 12(24):4020, 2024.
[4] Sven Goluža and Tomislav Kovačević and Stjepan Begušić and Zvonko Kostanjčar. Robot see, robot do: Imitation reward for noisy financial environments. In 2024 IEEE International Conference on Big Data (BigData), pages 4884–4891, 2024.
[5] Dukascopy Bank SA. Forex historical data feed, 2025.