Reinforcement Learning in Digital Finance

Explainable reinforcement learning

W.J.A. van Heeswijk
University of Twente
5.2.2025

Funded by
the European Union

Today’s agenda

- Link to XAI in (un)supervised learning
- Basics and framework of XRL
- Examples of XRL
- XRL in finance

Link to XAI in (un)supervised learning

Audience for explainable AI

- Explainable AI is a rapidly emerging field
  - Great need to open the black box and comprehend decisions made based on machine learning

- XAI may cater to various audiences, e.g.,:
  - Experts
  - Developers
  - End users
  - Executives

Refresher on XAI in (un)supervised learning [1/2]

- Model-Agnostic Methods: Work with any model type, offering flexibility.
  - **LIME** (Local Interpretable Model-agnostic Explanations): Generates local explanations by approximating models with simpler interpretable ones, focusing on feature importance for individual predictions.
  - **SHAP** (SHapley Additive exPlanations): Based on Shapley values from cooperative game theory, offering consistent, fair feature importance by attributing contributions across features for a prediction.

DIGITAL

Refresher on XAI in (un)supervised learning [2/2]

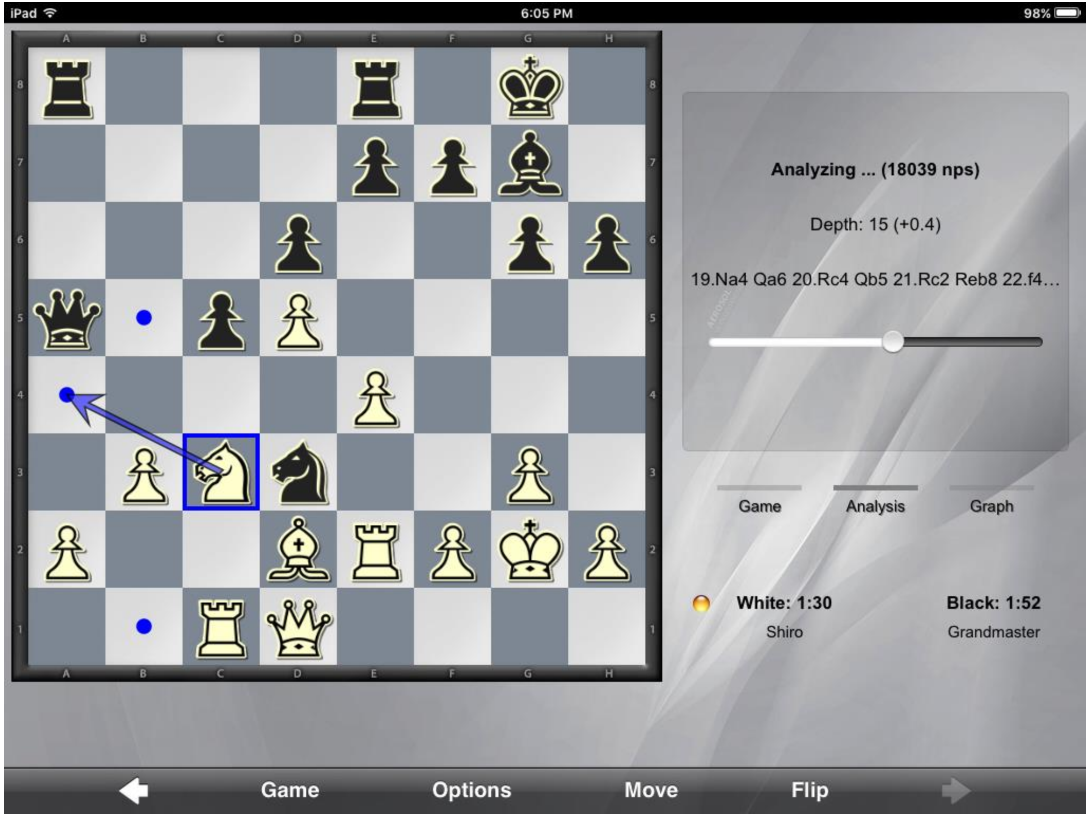
- Model-Specific Methods: Designed for specific model architectures (e.g., neural networks).
  - **Saliency Maps**: Highlights important regions in input (e.g., image pixels) that influence predictions.
  - **Feature Attribution**: Methods like integrated gradients calculate the contribution of each feature toward the prediction.

Reinforcement learning vs. supervised learning

- Supervised learning: predicting labels based on features
- Reinforcement learning: optimizing long-term reward by interactions

- Explaining predictions vs. actions
  - *"Why does the model take this specific action?"*
  - Requires insight into expected long-term rewards
  - Expected long-term rewards is captured in, e.g., deep Q network
  - Ambiguous how present-state features contribute to long-term rewards

Explainable?



Analyzing ... (18039 nps)

Depth: 15 (+0.4)

19.Na4 Qa6 20.Rc4 Qb5 21.Rc2 Reb8 22.f4…

White: 1:30
Shiro

Black: 1:52
Grandmaster

Source: https://commons.wikimedia.org/wiki/File:SmallFish_screenshot_%28Human_to_move,_computer_analysis_shown%29.png

What makes XRL different? [1/2]

- **Nature of feedback**
  - Feedback from the environment if often sparse, delayed, and depends on a policy that is updated over time
- **Decision-making process**
  - The impact of actions depends on complex stochastic factors such as transitions, rewards and possibly the policy itself
- **Temporal aspect**
  - Actions impact paths and cannot be viewed in isolation. It is the sequence of actions that yields certain long-term rewards

What makes XRL different? [2/2]

- **Exploration**
  - Both during learning and deployment, RL policies may exhibit inherent randomness, yielding suboptimal actions
- **System complexity**
  - The presence of multiple components such as policy, state transition and value function make it hard to pinpoint cause of results
- **Types of explanations**
  - Explanations require understanding of reward structure, value functions and state-dependent policy behavior

DIGITAL

Explainable Reinforcement Learning (XRL) framework

Directly interpretable models

- Directly interpretable RL models exist...
  - Linear policy
    - Feature weights are interpretable
  - Decision tree policy
    - Comprehensive decision paths

- ...however, most RL is 'deep' nowadays!
  - Neural networks perform very complex non-linear transformations
  - Generally considered non-interpretable

Framework for Explainable Reinforcement Learning

	Local	Global
Inherent		
Post-hoc		

Global vs. Local

- Global
  - Explain entire model
- Local
  - Explain individual actions

Global models

- Global models
  - Explain the general model behavior
  - Explain the overall model logic though its decision-making structures
  - Help users to trust the model

Local models

- Local models
  - Offer explanations for a specific decision
  - Explain why the model makes a certain decision in a given state
  - Identify the contributions of individual features on selecting the action
  - Help users to trust individual predictions

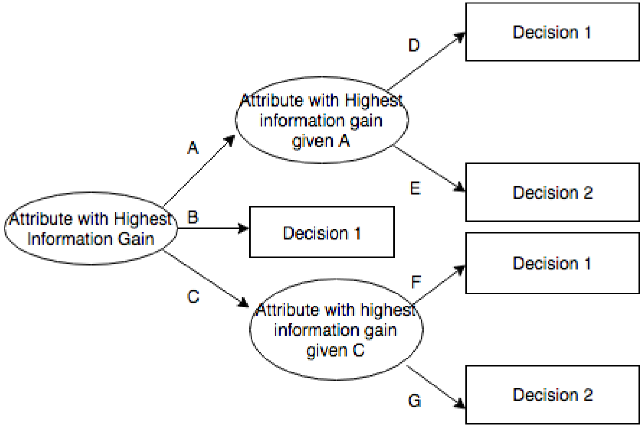
Inherent vs. Post-hoc

- Intrinsic model
  - Inherently interpretable during time of training

- Post-hoc model
  - Fit onto a more complex original model

Inherent models

- Intrinsic models
    - Restricts model complexity to be inherently interpretable at the time of training.
    - Tradeoff between transparency and accuracy, simpler models typically perform worse than complex ones.
    - Typically designed for one specific task


Attribute with Highest Information Gain
A
B
C
Attribute with Highest information gain given A
Decision 1
Attribute with highest information gain given C
D
E
F
G
Decision 1
Decision 2
Decision 1
Decision 2

Post-hoc models

- Post-hoc models
  - Analyzes the original model after training by fitting a simpler model that provides explanations for the original model
  - Preserves accuracy of the original (complex) model, at the cost of transparency
  - Can typically be applied to a variety of original models

Examples of XRL

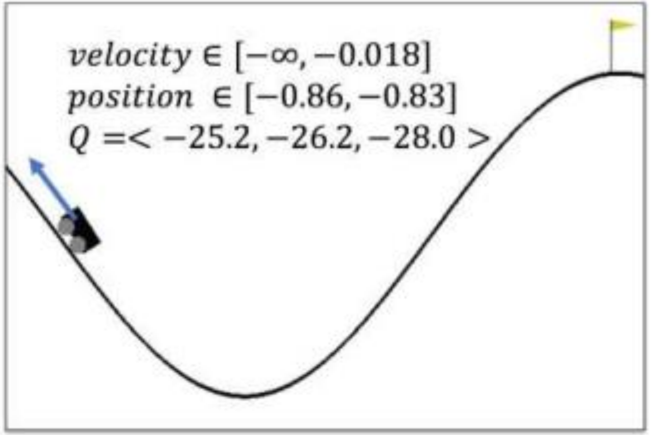
Explanation type

- Several formats to explain RL decisions and policies
  - Images
  - Text
  - Diagrams

- Typically, some sort of visual explanation is employed in XRL
- Format should match background of the audience
  - Many XRL formats require substantial expertise to interpret

Rule extraction

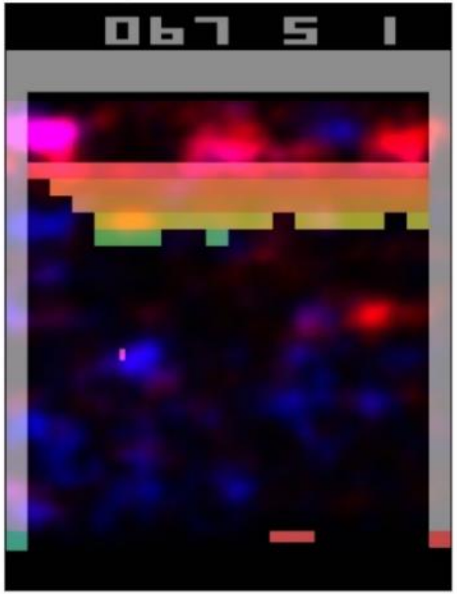
- Rule extraction maps policy to comprehensive rules
  - Highest average Q-value in range of state elements/features indicates which action to take


$velocity \in [-\infty, -0.018]$
$position \in [-0.86, -0.83]$
$Q =< -25.2, -26.2, -28.0 >$

Source: https://arxiv.org/pdf/1711.00138

Saliency maps

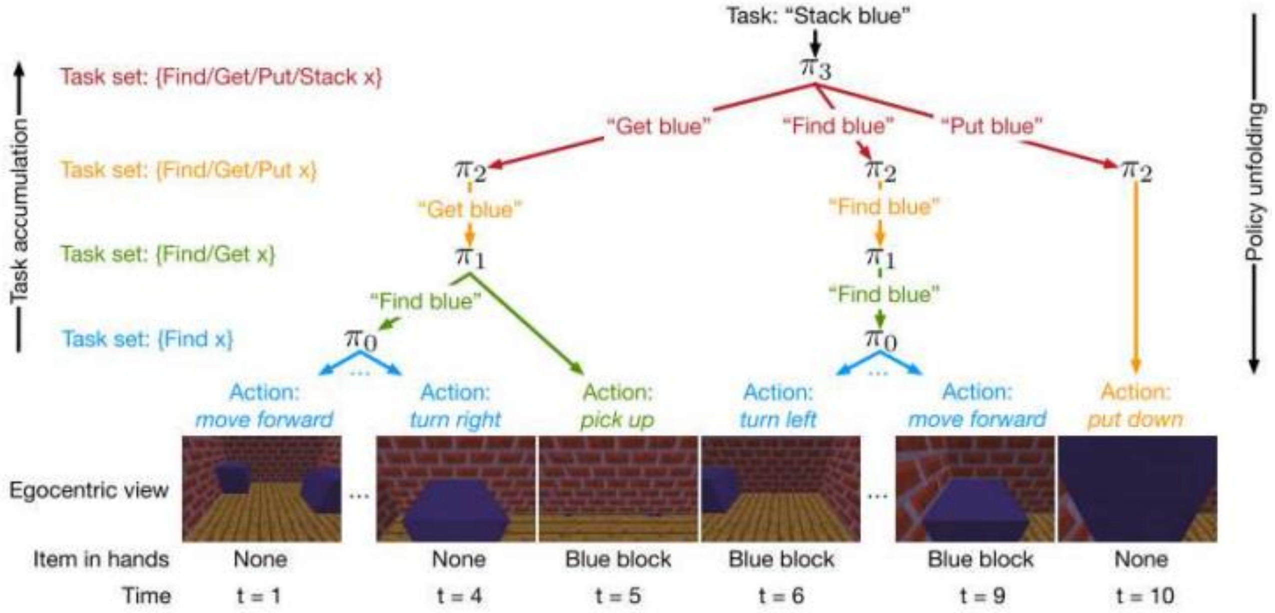
- Saliency map indicates high-value areas of the image
  - In picture: blue is actor saliency, red is critic saliency



Source: https://arxiv.org/pdf/1711.00138

Multi-level hierarchical policy

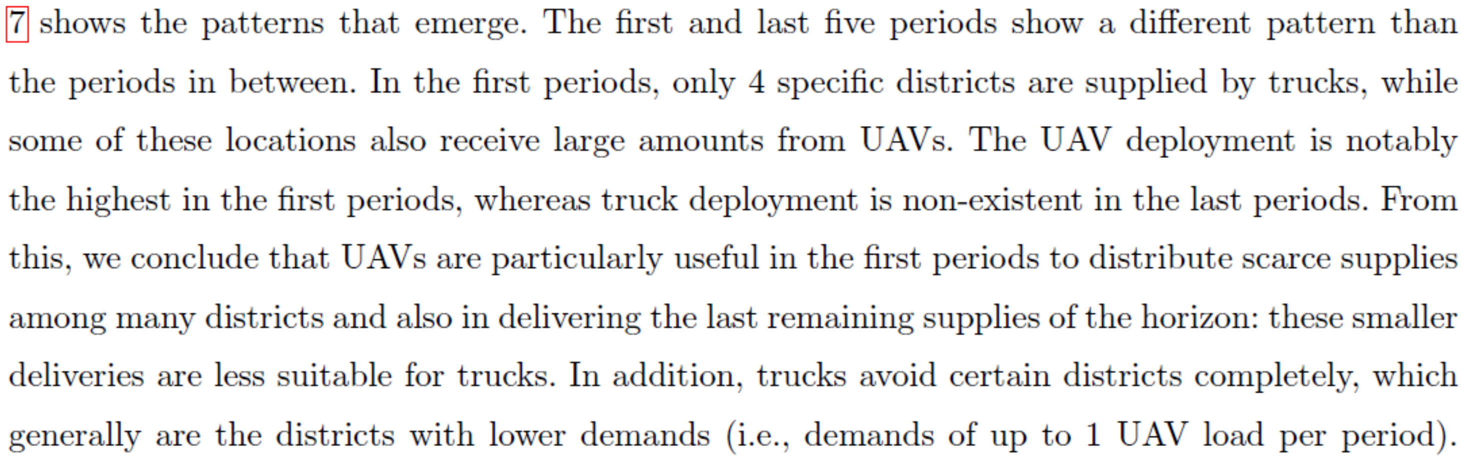
- Multi-level hierarchical policy sequentially describes actions
  - Relies on simple and comprehensive sub-policies



Source: https://arxiv.org/abs/1712.07294

Textual explanation

- Observations of decision characteristics and their impact on rewards explained in words

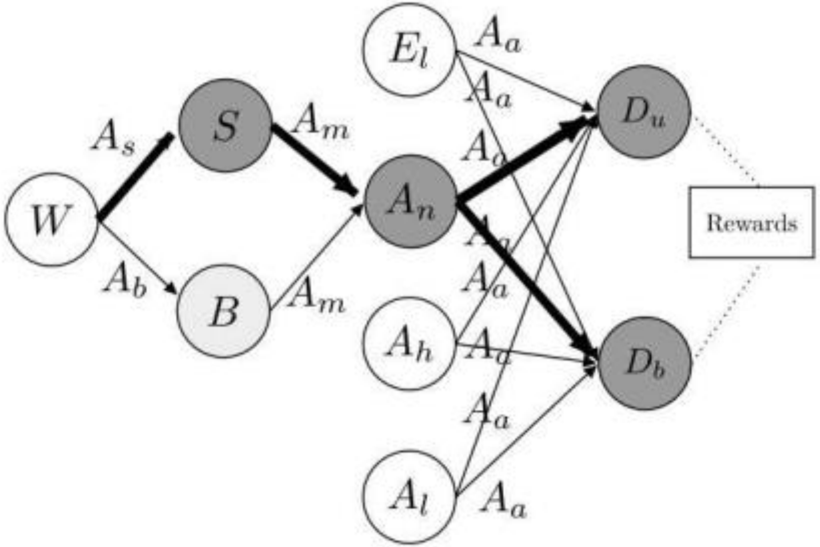


7 shows the patterns that emerge. The first and last five periods show a different pattern than the periods in between. In the first periods, only 4 specific districts are supplied by trucks, while some of these locations also receive large amounts from UAVs. The UAV deployment is notably the highest in the first periods, whereas truck deployment is non-existent in the last periods. From this, we conclude that UAVs are particularly useful in the first periods to distribute scarce supplies among many districts and also in delivering the last remaining supplies of the horizon: these smaller deliveries are less suitable for trucks. In addition, trucks avoid certain districts completely, which generally are the districts with lower demands (i.e., demands of up to 1 UAV load per period).

Source: https://arxiv.org/abs/2312.00140

Diagrams

- Factual and counterfactual explanation
    - Path of reasoning: “why this action, why not that action?

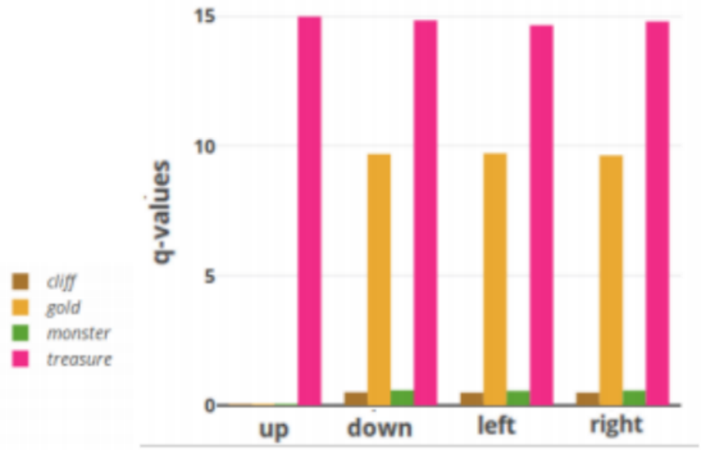


State variables:
W - Worker number
S - Supply depot number
B - barracks number
E - enemay location
A_n - Ally unit number
A_h - Ally unit health
A_l - Ally unit location
D_u - Destoryed units
D_b - Destroyed buildings
Actions:
A_s - build supply depot
A_b - build barracks
A_m - train offensive unit
A_a - attack

Source: https://ojs.aaai.org/index.php/AAAI/article/view/5631

Reward decomposition

- Learn Q-values for individual reward components
    - Allow to better understand why a certain action is selected

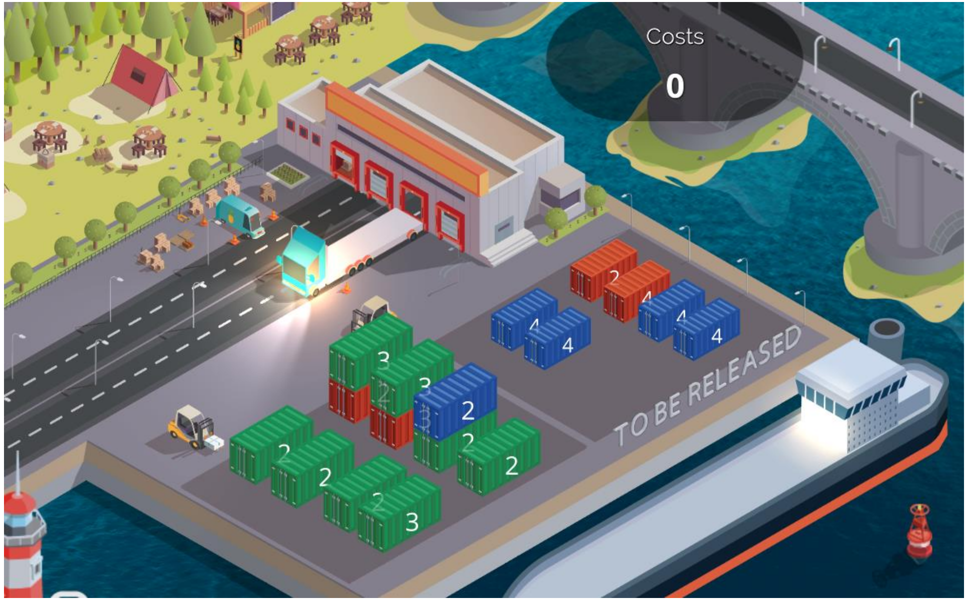


Source: http://web.engr.oregonstate.edu/~afern/papers/reward_decomposition__workshop_final.pdf

DIGITAL

Representation learning

- Learn abstract features that characterize (high-dimensional) state
  - Adds to comprehension of interaction between state, action and environment



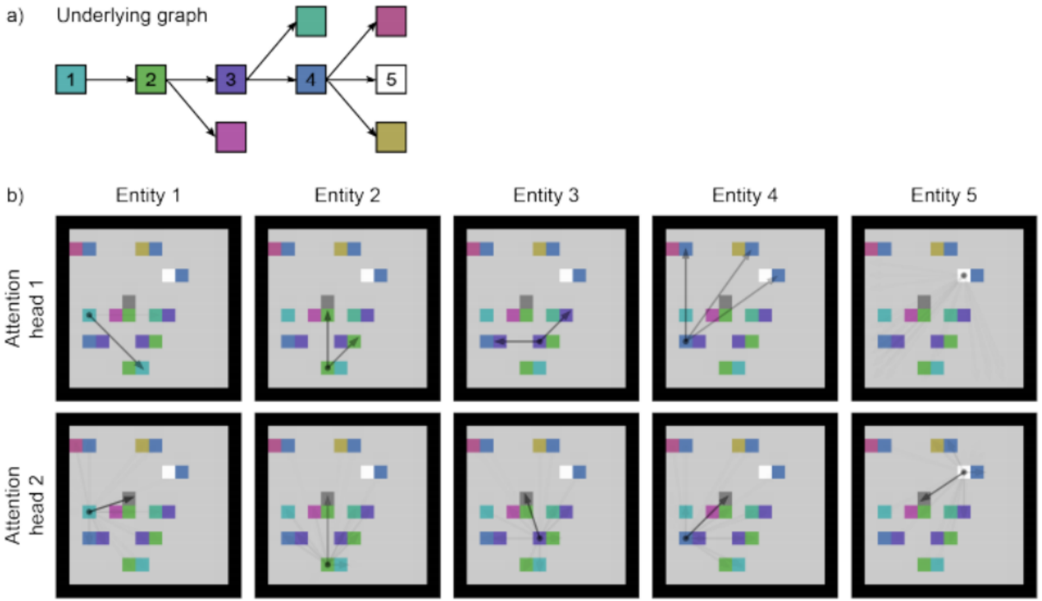
Many unique combinations of destination (color) and due date (number)

Potential features:
# red containers
# number of urgent containers

Opportunity to fit, e.g., linear regression model on features to derive Q-values

Representation learning – Another example

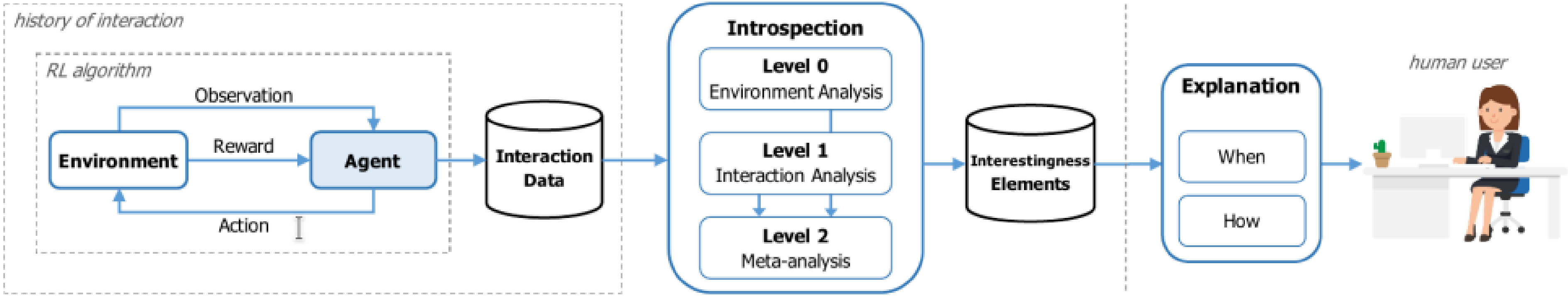
- Visualization of self-attention weights on a solution path



Source: https://arxiv.org/abs/1806.01830

Interestingness framework

- Extract interesting elements of interaction data – derived from statistical analysis – between agent and environment
  - Create appropriate explanation (e.g., a video) based on interesting elements



XRL in finance

From games to real world

- Vast majority of XRL solutions are applied to games
    - Clear structures, rules, and objectives
    - Strong reliance on visual interpretations
    - Often handcrafted solutions

- Real-world applications of XRL are still very much a green field
    - Many research opportunities
    - Necessity for real-world deployment of RL in finance!

Financial environment for XRL

- Constant dynamics and uncertainty → RL may adjust over time
  - Fluctuating stock prices
  - Economic developments
  - Changes in consumer behavior

- Periodic decisions to maximize rewards, e.g.,
  - Rebalance stock portfolio [why use which weights]
  - Decline loan requests [why decline?]
  - Flag transaction as potential fraud [why flag?]

XRL in Finance [1/3]

- Paper: “*Explainable Post hoc Portfolio Management Financial Policy of a Deep Reinforcement Learning agent*”
- Approach: compute SHAP and LIME values for portfolio weights

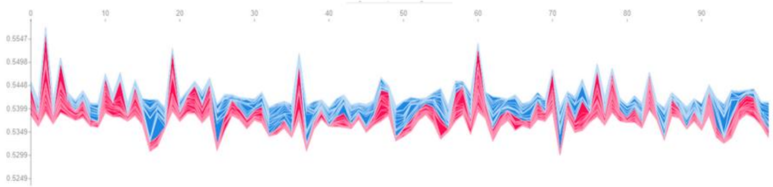


Fig. 5. SHAP force plot for AAPL weight allocation with all features contribution.

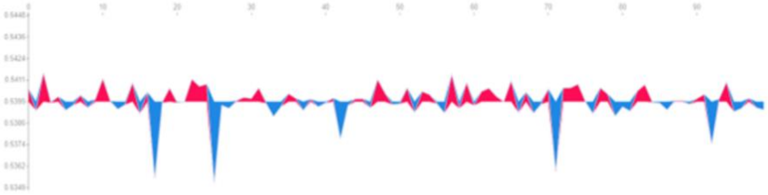
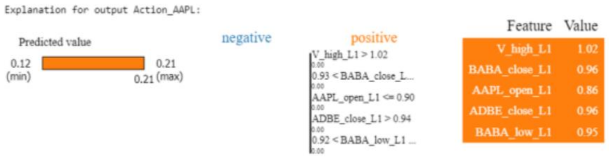


Fig. 6. SHAP force plot for AAPL weight allocation with only Apple’s closing price value contribution.


Explanation for output Action_AAPL:

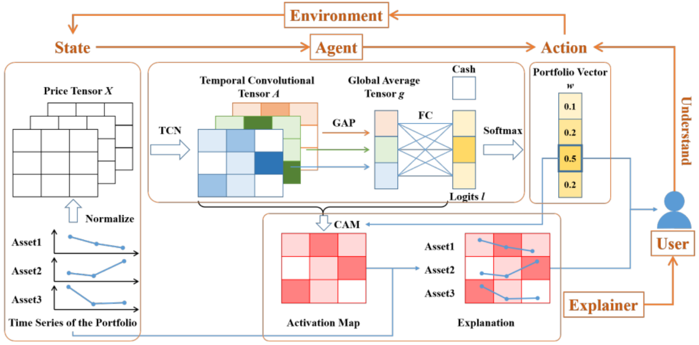
Feature	Value
V_high_L1	1.02
BABA_close_L1	0.96
AAPL_open_L1	0.86
ADBE_close_L1	0.96
BABA_low_L1	0.95

Fig. 7. LIME explanation for Apple’s weight allocation prediction at a particular instance.

Source: https://arxiv.org/pdf/2407.14486

XRL in Finance [2/3]

- Paper: “*XPM: An Explainable Deep Reinforcement Learning Framework for Portfolio Management*”
- Create activation map for an asset of interest.
  - Highlights the important assets and time intervals in the input state.
  - Explains which price movements drive decision to invest in asset.



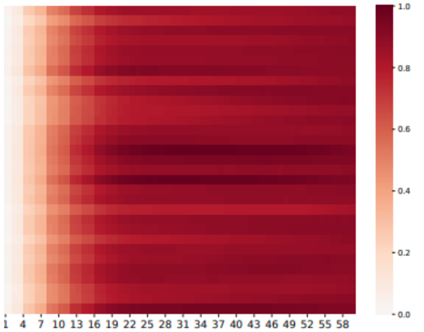
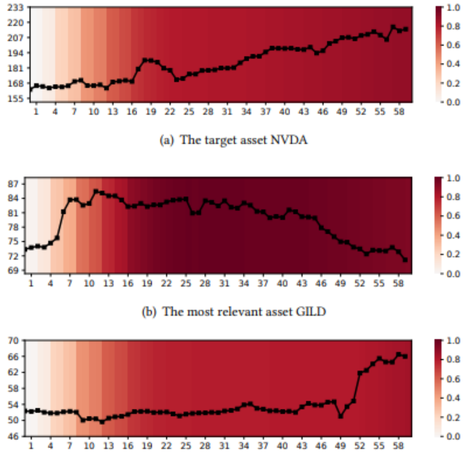


Figure 6: Activation map for NVDA in NASDAQ.



(a) The target asset NVDA

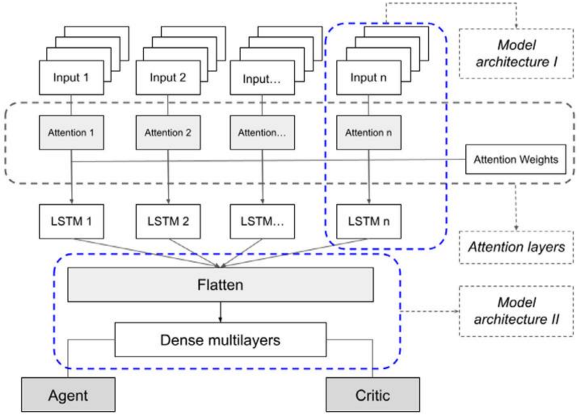
(b) The most relevant asset GILD

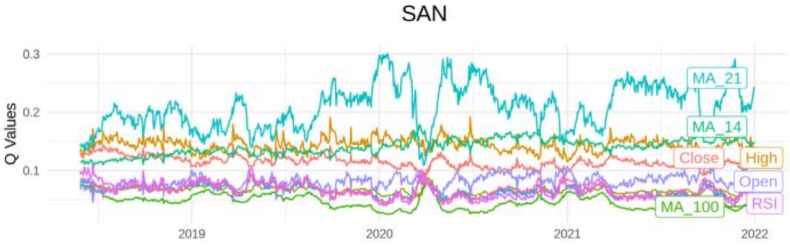
(c) The least relevant asset QCOM

Source: https://dl.acm.org/doi/pdf/10.1145/3459637.3482494

XRL in Finance [3/3]

- Paper: “*XPM: An Explainable Deep Reinforcement Learning Framework for Portfolio Management*”
- Independent attention layers in LSTM
  - Discover critical features per asset





Source: https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13667

XRL in Finance

- XRL in finance is still in its infancy
  - A few works on stock trading and portfolio optimization
  - No generally accepted solution method
  - Some explanations are quite advanced, aimed at ML experts
  - Many concepts from XRL still untested in finance!

Wrapping up

Closing words

- XRL is a recent field in development
  - XRL in finance even more so, very few applications
  - Remind that many decisions in finance must be explainable!
  - Plenty of practical challenges and research opportunities

Tutorial

- **Review problem description** and discuss:
  - What aspects of the RL model should be explained? (Policy, actions, reward function?)
  - Who needs the explanation? (Regulators, investors, end users?)
  - What techniques seem most appropriate?
- **Task:** Outline an explainability solution for their problem, including:
  - **Data & model:** What RL model is used? What data is involved?
  - **Explanation approach:** What XAI/XRL technique(s) will be applied, and why?
  - **User interaction:** How will explanations be presented (graphs, reports, interactive tools)?
- Each group gives a **2-minute summary** of their explainability solution

DIGITAL

Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe: Marie Skłodowska-Curie Actions. Neither the European Union nor the granting authority can be held responsible for them.

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101119635

DIGITAL