Symbolic Explanation of Affinity-Based Reinforcement Learning Agents with Markov Models

Charl Maree[a,b,*], Christian W. Omlin[a]

[a] *Center for Artificial Intelligence Research, University of Agder, Grimstad, 4879, Norway*
[b] *Chief Technology Office, Sparebank 1 SR-Bank, Stavanger, 4007, Norway*

Abstract

The proliferation of artificial intelligence is increasingly dependent on model understanding. Understanding demands both an interpretation—a human reasoning about a model's behavior—and an explanation—a symbolic representation of the functioning of the model. Notwithstanding the imperative of transparency for safety, trust, and acceptance, the opacity of state-of-the-art reinforcement learning algorithms conceals the rudiments of their learned strategies. We have developed a policy regularization method that asserts the global intrinsic affinities of learned strategies. These affinities provide a means of reasoning about a policy's behavior, thus making it inherently interpretable. We have demonstrated our method in personalized prosperity management where individuals' spending behavior in time dictate their investment strategies, i.e. distinct spending personalities may have dissimilar associations with different investment classes. We now explain our model by reproducing the underlying prototypical policies with discretized Markov models. These global surrogates are symbolic representations of the prototypical policies.

Keywords: Explainable AI, personalized financial services, policy regularization, affinity-based learning, Markov models

*Corresponding author. Email address: charl.maree@uia.no

1. Introduction

The ultimate goal of explainable AI is understanding. It builds trust, improves safety, and improves predictive performance by facilitating precise model improvements [1]. For instance, feature saliency analyses can improve feature selection and consequently the predictive performance in stock trading [2], and rule extraction can enhance trust in an AI system for loan approvals Sachan et al. [3]. Despite considerable advancement in fields such as explainable reinforcement learning (RL) [4, 5, 6], the explainability of their underlying models has not yet been fully addressed [7, 8].

Reinforcement learning has become omnipotent in finance, for example, multi-agent RL for algorithmic trading [9]. Methods such as probabilistic argumentation [10], structural causal modeling [11], and introspection through interesting elements [12] exemplify the pursuit of post-hoc explainability. We, however, propose an alternative approach: rather than attempting to extract the learned strategy post hoc, ours is an intrinsic method that instills a desirable behavior during training [13]. Through regularization of the objective function, our method encourages global action affinities and thus exercises control over what agents learn. We have demonstrated the value of our method in personal prosperity management, where individual spending behaviors dictate investment strategies [14]. We instilled affinities for certain asset classes into the policies of a set of prototypical agents, each associating with a given personality trait. For example, a conscientiousness agent prefers asset classes typically associated with reduced risk.

Understanding ensues from a model explanation and an interpretation of its behavior. We distinguish between these two concepts: an explanation is a symbolic representation of a model's predictions, while an interpretation is a human reasoning about its behavior. While our agent's policies are inherently interpretable, they lacked a symbolic explanation. Using discretized Markov models, we now provide that explanation and thus gain insight into previously unanswered questions, such as why do the agents invest according to conventional wisdom: exploiting the benefits of compound growth and reducing risk with increasing customer age. These previously unanswered questions demonstrate the need for both explanations and interpretations: the lack of a symbolic representation of agents' policies inhibited our complete understanding.

Our contributions are therefore: (1) we demonstrate how to instill global action affinities, thus affecting how RL agents learn, which we argue is a useful

paradigm shift over the current approach of either post hoc rule extraction or constrained learning, (2) we distinguish between model explainability and interpretability, and in an empirical example demonstrate the difference and the utility of both, and (3) we propose a method of using Markov models to extract symbolic explanations of RL agents' policies. In the next section, we provide an overview of the current state of the art in explainable RL and identify limitations in the field. We then describe our data and empirical methodology, discuss our results, and conclude with insights and future work.

2. Related Work

RL agents learn to solve problems by maximizing the total expected reward awarded by the environment in which they act. They are particularly adept at learning in the presence of sparse and delayed rewards [15]. The environment is a discrete-time process where the current state depends only on the previous state and the action taken by the agent: a Markov decision process (MDP), described by the tuple (S, A, R, P) where S is a set of states, A a set of actions, $R(s, a)$ the reward for taking action $a \in A$ in the state $s \in S$, and $P(s, a) = P(s'|s, a)$ the probability that action a in the state s leads to the state s' [16]. Deep deterministic policy gradients (DDPG) is a model-free RL algorithm for learning policies in a continuous action space [17]. A DDPG agent consists of four neural networks: an actor $\mu(\theta)$ representing the policy, a critic $Q(\theta)$ representing the state action value function, and for numerical stability, a target actor $\mu'(\theta')$ and a target critic $Q'(\theta')$. During learning, the target network parameters are typically updated slowly given a soft update parameter $\tau \in [0, 1]$ with a small value: $\theta'_i = \tau\theta_i + (1 - \tau)\theta'_i,\ i \in \{\mu, Q\}$.

Explainable RL has traditionally employed generic methods that explain the underlying models of agents [1]. More recently, however, bespoke methods have emerged that consider the state-action space and / or the behavior of the learned policy [6, 4, 5]. Most, if not all, of these approaches extract explanations after training; they generalize the learned policy through observation or statistical analyses. Few of these extracted explanations match our definition of explainability, and most are more accurately described as interpretations. *State representation learning* connects the state space with information from actions, rewards, or expert knowledge when extracting representations that are useful for reasoning about policies [18]. Under certain restrictions, e.g., linearity, it learns models that either predict states form state-action pairs, or actions from states, thus simplifying the state-action

space and improving interpretability. *Introspection* analyzes an agent's experience through statistics such as the frequency of occurrences of states, state-actions, and transitions, the transition probabilities, and estimated rewards compared to the learned state-action value function [12]. It uses interesting elements from this analysis, such as outliers, mean values, etc. to reason about agents' behaviors. *Structural causal modeling* learns causal relationships between states, actions, and rewards by defining action influence graphs that map the action transitions for all possible paths from an initial state to a set of terminal states [11]. It defines the causal chain as the one path in the action influence graph that matches the learned policy, and a reward chain as the vector of rewards along this causal chain. Its interpretation of the policy is the comparison between the reward chain and all other possible reward vectors that do not follow the causal chain. *Probabilistic argumentation* uses argumentation graphs—sets of attacking and supporting arguments for each action in a finite action space—to learn interpretations in a RL setting [10]. The state is the intersection of the argumentation graph and the policy to be explained, the actions form a probabilistic distribution across the arguments, and the rewards depend on whether an argument attacks or supports the current action. The learned policy provides probabilistic interpretations of agents' actions in human understandable terms: supporting and attacking arguments for each action. *Reward decomposition* replaces the scalar reward with a vector of more meaningful rewards, where the total reward is the sum of the vector [19, 20, 21]. Although evaluating the reward vector for a given action might enable reasoning about that action in meaningful terms, it does not take into account expected future rewards and can be insufficient in environments with delayed or sparse rewards. *Reward redistribution* addresses this problem by redistributing delayed rewards in time; it assigns credit to previous actions, thus reducing the delay of the reward [22]. The immediate reward for each time step in a sequence of state-action transitions is equal to the change in the total expected reward. It defines key interpretable events in the policy and, through sequence alignment, redistributes rewards to those events given a set of transition sequences. *Hierarchical RL* divides complex tasks into smaller and simpler tasks that are solved by correspondingly simpler RL agents [23, 21]. An orchestration agent learns to sequentially combine these prototypical agents to solve complex tasks. If tasks are sufficiently subdivided, the interpretation, or human reasoning about agents' decisions, follows from their simplicity.

The complexity of RL models exacerbates the issue of fidelity and vali-

dation of any post hoc explanation. We, instead, encourage agents to adapt their behavior during learning, thus instilling an inherent probabilistic action affinity that is also an interpretation of their behavior [13]. Contrary to constrained RL, which avoids certain conditions [24, 25], affinity-based learning promotes certain behaviors. This paradigm shift allows the developer to define a desired behavior that an agent must follow during learning, thus instilling a characterization and interpretation during learning; it decouples learned strategies from the reward expectation [26]. Affinity-based RL is not to be confused with preference-based RL that completely eliminates the reward function and instead learns state-action trajectories that maximize the preferences of the expert between pairs of state-action combinations [27]. Affinity-based RL uses policy regularization that aids—and is never detrimental to—learning convergence by encouraging exploration in environments with complex dynamics or particularly sparse rewards [28, 29]. It adds a term to the objective function that penalizes any divergence between the current policy and a given prior, for example, Kullback-Leibler (KL) regularization, which uses KL divergence as the distance measure [30]. Entropy regularization is a specific case of KL-regularization, where the prior is a uniform action distribution that increases the entropy of the policy and thus encourages general exploration of the state-action space[31]. Our method instead encourages exploration of a predefined subset of the state-action space, which describes the desired behavior [13]. We define our objective function as follows:

$$J(\theta) = \mathbb{E}_{s,a\sim\mathcal{D}}\left[R(s,a)\right] - \lambda L \tag{1}$$
$$L = \frac{1}{M}\sum_{j=0}^{M}\left[\mathbb{E}_{a\sim\pi_\theta}(a_j) - (a_j|\pi_0(a))\right]^2$$

where $\mathcal{D}$ is the replay buffer, λ is a hyperparameter that scales the regularization term L, M is the number of actions, and π_0 is a specific prior action distribution that represents the desired behavior. Instilling an interpretable behavior is sufficient for online policy interpretation [32]. Unlike KL-regularization, our prior π_0 is independent of the state and therefore instills a global action affinity in the learned policy. We have demonstrated this in Maree and Omlin [13] where agents navigated a grid towards a destination; they learned to prefer, for example, only right turns and followed optimal paths given their global affinities. In a more elaborate example, we trained a set of prototypical agents with global affinities to invest in certain asset

classes [14]. We observed the emergence of interesting investment strategies, such as capitalizing on compound growth and reducing risk with portfolio maturity. Although consistent with conventional wisdom, these strategies were absent from the objective function. To complete our understanding of this behavior, we now provide a symbolic representation—an explanation—of these policies using Markov models.

A hidden Markov model (HMM) models an unobservable Markov process X from its relation to an observable Markov process Y; it learns about X by observing Y under the key assumptions that Y_t is solely dependent on X_t, and X_t is solely dependent on X_{t-1} (the Markovian property) [33]. For a finite hidden state space X, there exists a Markov matrix F—the sum of the rows add up to one—of state transition probabilities where $F_{ij} = P(X_{n+1} = j \mid X_n = i)$. Similarly, for a finite observed state space Y, there exists a Markov matrix E that describes emission probabilities: $E_{ij} = P(Y_t = j \mid X_t = i)$. We illustrate this process in Figure 1. Given a series of observed states $\{Y_t\}_{t=0}^{T}$, the transition and emission probabilities can be estimated using the Baum-Welch algorithm—a special case of the expectation-maximization algorithm [34].

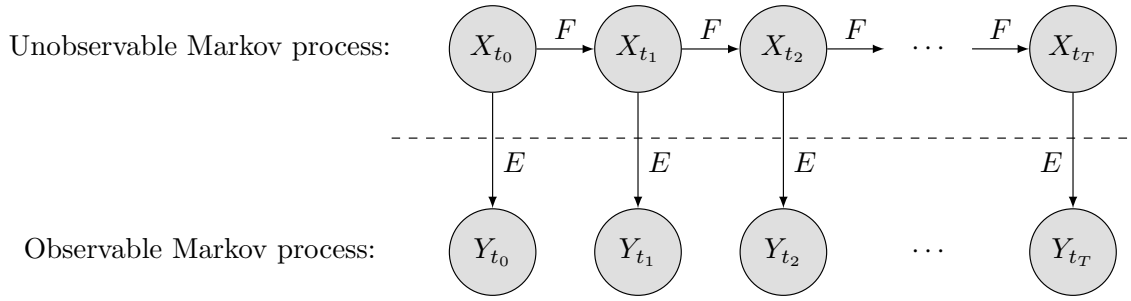

Figure 1: A trellis diagram representing a hidden Markov model with an unobservable Markov process X, and observable Markov process Y, transition probability matrix F, and emission probability matrix E.

3. Methodology

In Maree and Omlin [35], we defined a set of prototypical agents with intrinsic investment behaviors associated with each of five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. We used affinity-based RL to learn investment strategies for each of the prototypical agents. Their actions were monthly investment distributions across

five different asset classes: savings accounts, property funds, stocks, mortgage curtailment, and luxury items. While stocks, savings, and property investments are self-explanatory, we defined mortgage curtailment as additional payments that reduce the principal debt of a loan, and luxury items such as art, classic cars, fine wines, etc., that might appear in, e.g., the Knight Frank luxury investment index [36]. We also learned linear combinations of these agents to best match the spending personalities of individual customers which, for the sake of brevity, we do not discuss here. However, to facilitate an understanding of our application, we summarize this paradigm in Figure 2 and refer the reader to a comprehensive account in [35]. We now provide an explanation for the prototypical agents' policies using Markov models.

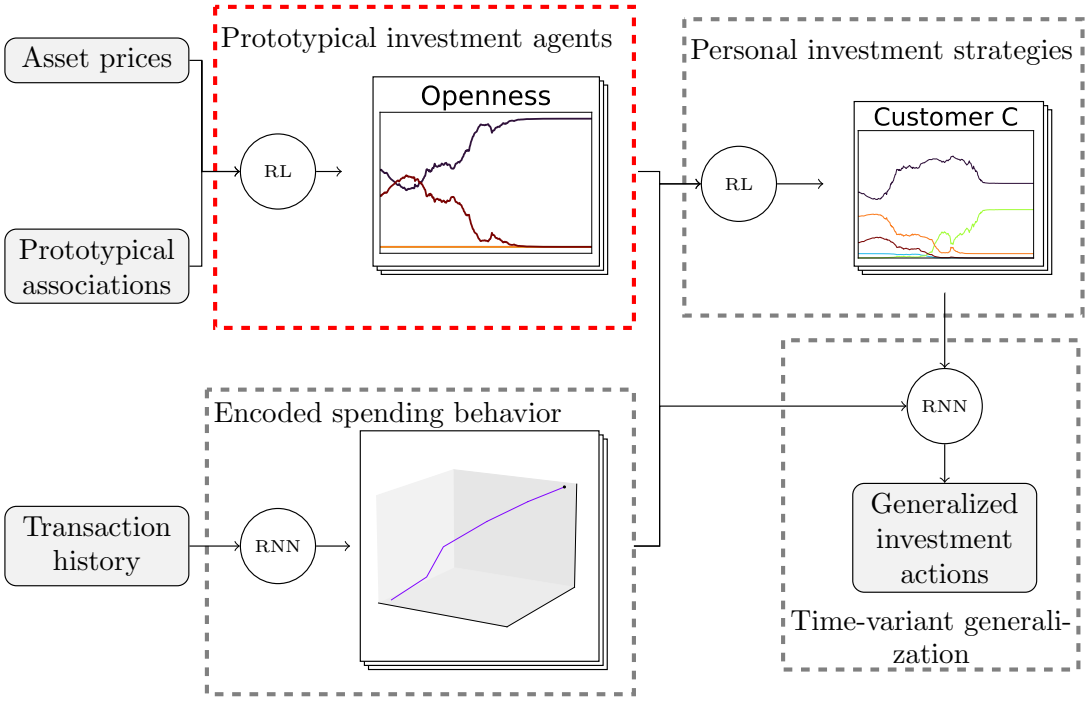

Figure 2: A flow diagram illustrating our system of RL agents that predict personalized investment strategies. There are five prototypical affinity-based RL agents (enclosed in a red dashed rectangle), each associating with one of five personality traits: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These are the agents that we explain using Markov models. Their actions are combined to match the spending behaviors of individual customers, and these combinations are continuously adjusted according to their changing spending behavior using a recurrent neural network (RNN). While these combinations are outside of the scope of this study, we believe it is useful to illustrate how the agents are used in a complete application.

To train our agents, we used pricing data for the S&P500 index, Norwegian property index, and the Norwegian interest rate index between 1994 and 2022. We used two common market indicators—the moving average convergence divergence (MACD) and the relative strength index (RSI) [37]—to capture market dynamics. These indicators are the state space features of the environment in which our agents learned. We show these features in Figure 3. There is an additional state variable that indicates the maturity of the portfolio; its value is 0.0 in the first month (January 1994) and linearly increases to 1.0 in the final month (December 2021).

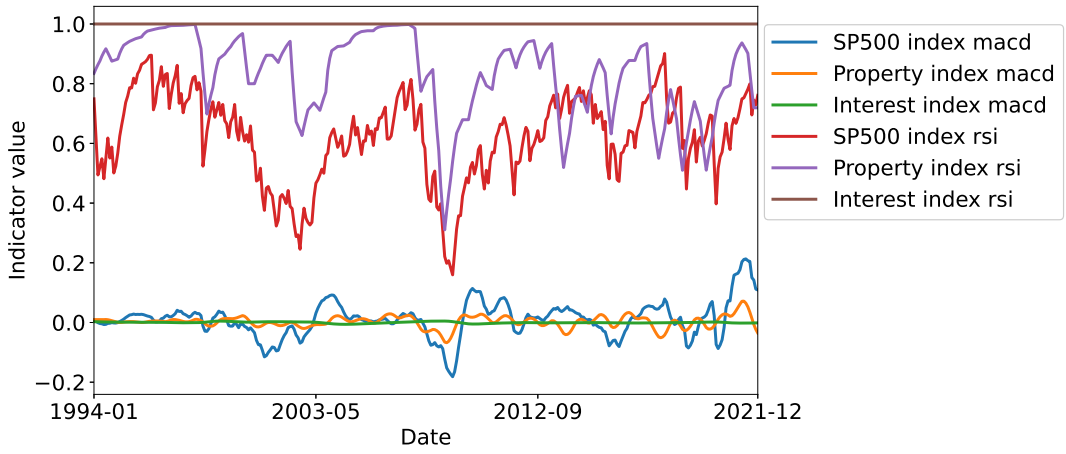

Figure 3: The state data used to train the prototypical agents. We used two common market indicators—the moving average convergence divergence (MACD) and relative strength index (RSI)—to represent market dynamics of the S&P500 index, Norwegian property index, and Norwegian interest rate index. Our learning time frame was between 1994 and 2022.

We show the resulting policies for the five prototypical agents in Figure 4. The agents optimized a common reward function, i.e., monthly returns; they maximized the portfolio value. Though they shared a common reward function, the agents learned unique investment strategies: the conscientiousness agent, for instance, prefers low-risk investment in property followed by resolute mortgage curtailment, while the openness agent prefers investments that might incite their curiosity, such as luxury items and stocks.

To train Markov models that match the predictions of the five prototypical agents, we discretized the states and actions of the agents. We assigned three bins to the RSI indicator based on the knowledge that values between 0 and 0.3 indicate oversold conditions, values between 0.7 and 1 indicate over-

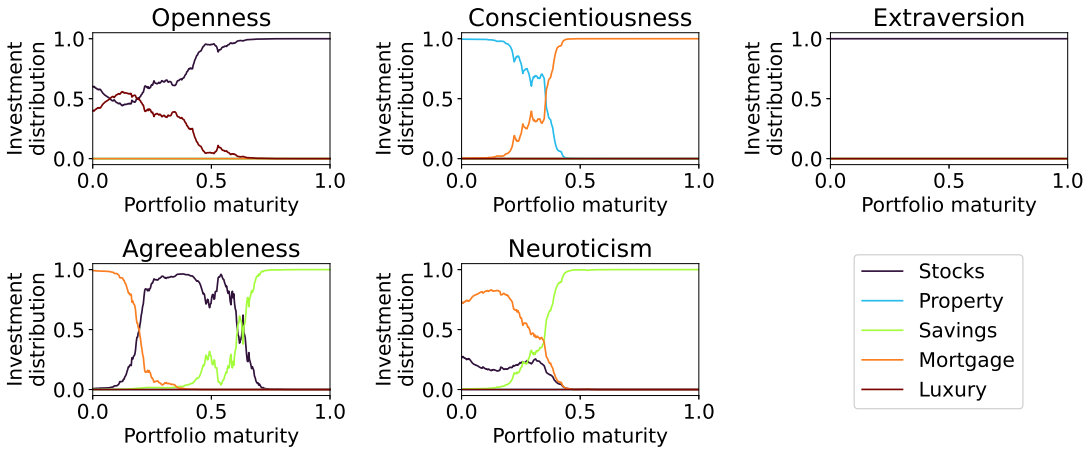

Figure 4: The monthly actions of the five prototypical agents, shown on an x-axis ranging between 0 to 1, representing months between 1994 and 2022. The y-axis represents the monthly investment in each of the asset classes. Note that the actions strictly represent the purchase of assets, i.e. the extraversion agent, for instance, consistently invests 100% of available monthly funds into stocks, thus consistently increasing the portfolio holding of stocks; assets are never sold. Though the agents optimised a common reward function—monthly returns—, their distinct strategies were instilled through affinity-based learning.

bought conditions, and values between 0.3 and 0.7 are inconclusive [37]. We similarly assigned two bins for the MACD indicator based on the knowledge that positive values represent a buy signal, while negative values represent a sell signal [37]. We divided the maturity state feature into 28 bins: one for each year of the investment period. We finally assigned 5 equally sized bins for the agents actions, between 0 and 1. This resulted in 168 potential states, of which only 102 states ever occurred. It is reasonable that not all possible states occurred, since MACD and RSI are related; it is not unexpected that whenever RSI indicates oversold conditions, MACD could suggest a buy signal [37]. We then estimated the transition probabilities in the Markov matrices F_i and the emission probabilities E_i, where $i \in [1, 5]$, for the five Markov models by observing the state transitions and the corresponding actions for each of the prototypical agents. Using the initial state and the five Markov models defined by F_i and E_i, we can reproduce the policies of the five prototypical agents with high fidelity.

4. Results

We trained five distinct Markov models as global surrogates to reproduce the predictions of five affinity-based RL agents. We show the discretized actions of the agents and the corresponding predictions of the Markov models in Figure 5. Using only the initial state as input, the Markov models predict the agents' actions with high fidelity, with some uncertainly when action values change due to the probabilistic nature of Markov models.

Figure 6 shows the state transitions for a non-exhaustive subset of states: the first 16 states visited including the initial state. We observe that not all states are visited, which is expected since the market indicators MACD and RSI are not entirely independent, nor are the stock and property markets in general. For example, during macroeconomic downturns we often observe a decline in both these markets: refer to Figure 3 and observe, for example, the decline in both the property and S&P500 indices during the 2008 recession. Property and stock markets can also demonstrate an inverse correlation: in Figure 3 the RSI curves for property and stocks can have reversed slopes, while the MACD curve can exist on opposite sides of zero. By perturbing the sizes and number of bins, we observed that portfolio maturity holds the most salient information. This is an important observation; it suggests that the values of the market indicators have a lesser influence on investment strategies compared to the maturity of the portfolio. This is in line with conventional wisdom that long-term investment should not be overly concerned with short-term market volatility; property and stock indices have typically followed an upward trend in the long run. The reduced dependence on market conditions increases confidence in model robustness when trading on unseen data: the unseen market conditions are less important than investor age; the basic principle that younger investors can afford increased risk in return for higher reward, and mature investors should seek to reduce portfolio risk, is common across a wide range of market conditions.

5. Conclusions

Understanding deep AI models requires an interpretation of their behavior and a symbolic representation, or explanation, of their functioning. These two elements facilitate reasoning about a model and, thus, enhance trust in its decisions. We have proposed a novel affinity-based approach to interpretable reinforcement learning; it encourages exploration of a predefined

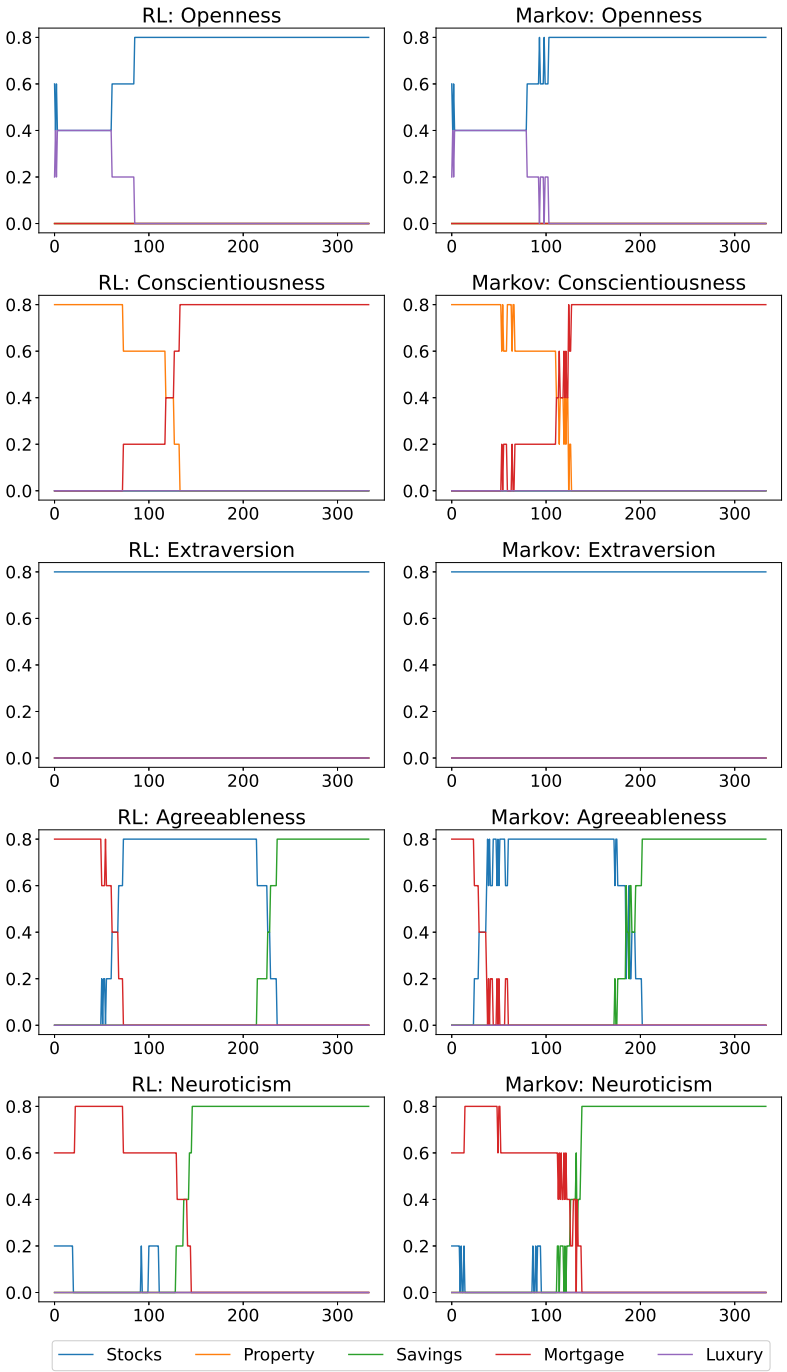

Figure 5: A visual comparison between the discretized predictions of five RL agents (on the left) and the five corresponding Markov models (on the right). The single input to the Markov models is the initial state, from which they predict the transition to the next state and the corresponding action by the agent. The Markov models clearly predict the actions with high fidelity.

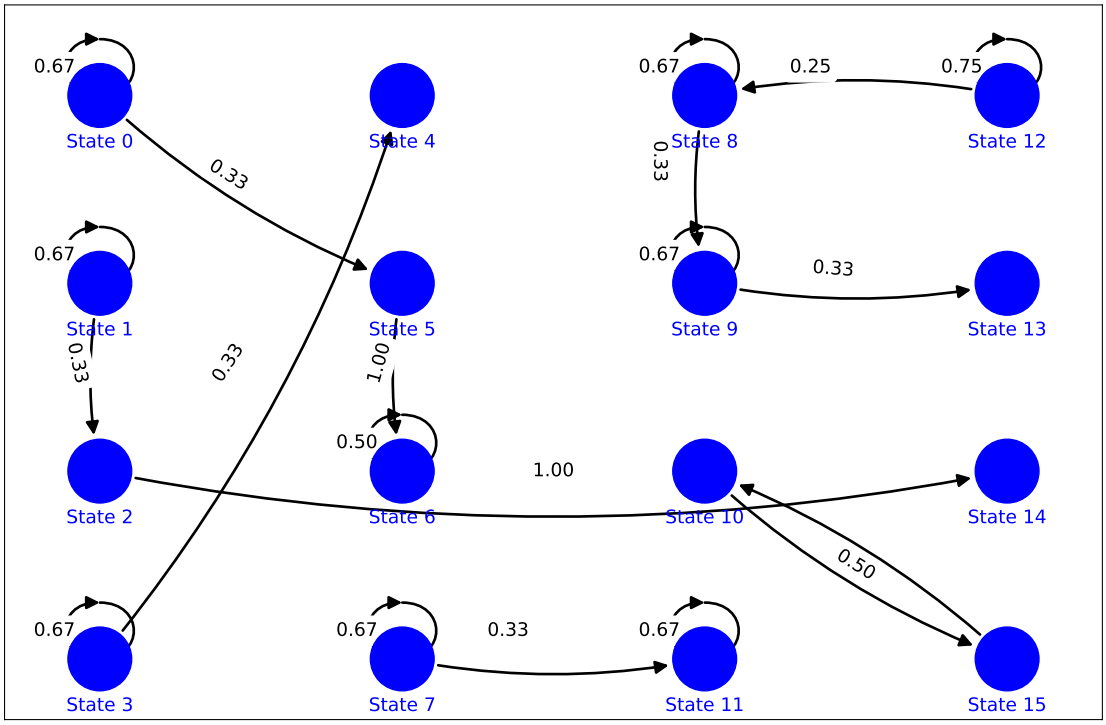

Figure 6: A non-exhaustive illustration of the trained Markov model showing state transitions for a subset of states. States are shown as blue circles, and state transitions and their probabilities are shown in black. We show the first 16 states, as visualizing all 102 states is not feasible. Each state represents a set of features with discretized values for MACD and RSI indicators of the property and stock indices, respectively, as well as the maturity of the portfolio. Note that not all state transitions are shown, since the origin or destination state might not be included in this subset.

subset of the state-action space. This prior action distribution describes the agent's desired behavior and is the interpretation of its policy. However, our solution lacked a symbolic explanation, resulting in unanswered questions about why they make certain decisions. A concrete example is why a set of agents, that learned to invest according to the preferences of prototypical personality traits, invest in more risky assets for younger investors and reduce risk with investor age. We now provide a symbolic representation of the agents' policies, using Markov models, that answer such questions. Our Markov models recreate, with high fidelity, the discretized investment strategies of five prototypical investment agents using only the initial state. By perturbing the bin sizes of the discretized state features, we are able to determine the most salient feature: portfolio maturity. The fact that

market conditions play a diminutive role in model prediction is significant: it enhances trust in out-of-sample predictions and suggests that investment timing is more important than market conditions. The agents make use of compounding growth by investing in higher reward—but more risky—assets early on, and fulfill their prescribed action distributions towards the end of the investment period; they learned how to maximize rewards. This use case demonstrates the need for both interpretations and explanations to fully comprehend the functioning and characterization of deep RL systems. The Markov model is a valuable tool for extracting a symbolic representation of an otherwise opaque RL model, and affinity-based RL is a unique approach to control what RL agents learn and thus interpret their behavior. It is a paradigm shift from current approaches that either encourage general exploration for the purpose of improved convergence or constrain the state space to prevent the policy from visiting undesirable states. It is compelling to apply affinity-based RL to virtuous agents, personalized learning and teaching, chronic disease treatment, climate change, wind farm operations, etc.

Declaration of competing interest

The authors declare that they have no competing interests.

Funding

This study was partially funded by a grant from the Norwegian Research Council, project number 311465.

References

[1] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[2] S. Carta, S. Consoli, A. S. Podda, D. R. Recupero, M. M. Stanciu, Statistical arbitrage powered by explainable artificial intelligence, Expert Systems with Applications 206 (2022) 117763.

[3] S. Sachan, J.-B. Yang, D.-L. Xu, D. E. Benavides, Y. Li, An explainable ai decision-support-system to automate loan underwriting, Expert Systems with Applications 144 (2020) 113100.

[4] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowledge-Based Systems 214 (2021) 1–24.

[5] L. Wells, T. Bednarz, Explainable AI and reinforcement learning: A systematic review of current approaches and trends, Frontiers in Artificial Intelligence 4 (2021) 1–48.

[6] E. Puiutta, E. M. Veith, Explainable reinforcement learning: A survey, Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science 12279 (2020) 77–95.

[7] Y. Ramon, R. Farrokhnia, S. C. Matz, D. Martens, Explainable AI for psychological profiling from behavioral data: An application to big five personality predictions from financial transaction records, Information 12 (2021) 1–28.

[8] L. Cao, Ai in finance: Challenges, techniques and opportunities, Banking & Insurance eJournal 14 (2021) 1–40.

[9] A. Shavandi, M. Khedmati, A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets, Expert Systems with Applications 208 (2022) 118124.

[10] R. Riveret, Y. Gao, G. Governatori, A. Rotolo, J. V. Pitt, G. Sartor, A probabilistic argumentation framework for reinforcement learning agents, Autonomous Agents and Multi-Agent Systems 33 (2019) 216–274.

[11] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, arXiv 1905.10958v2 (2019).

[12] P. Sequeira, E. Yeh, M. Gervasio, Interestingness elements for explainable reinforcement learning through introspection, Joint Proceedings of the ACM IUI Workshops 2327 (2019) 1–7.

[13] C. Maree, C. W. Omlin, Reinforcement learning your way: Agent characterization through policy regularization, AI 3 (2022) 250–259.

[14] C. Maree, C. W. Omlin, Can interpretable reinforcement learning manage prosperity your way?, AI 3 (2022) 526–537.

[15] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, second ed., The MIT Press, 2018.

[16] R. Bellman, A markovian decision process, Journal of mathematics and mechanics (1957) 679–684.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv 1509.02971 (2019).

[18] T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, D. Filliat, State representation learning for control: An overview, Neural Networks 108 (2018) 379–392.

[19] H. van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, J. Tsang, Hybrid reward architecture for reinforcement learning, arXiv 1706.04208 (2017).

[20] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, F. Doshi-Velez, Explainable reinforcement learning via reward decomposition, International Joint Conference on Artificial Intelligence. A Workshop on Explainable Artificial Intelligence. (2019).

[21] L. Marzari, A. Pore, D. Dall'Alba, G. Aragon-Camarasa, A. Farinelli, P. Fiorini, Towards hierarchical task decomposition using deep reinforcement learning for pick and place subtasks, arXiv 2102.04022 (2021).

[22] M.-C. Dinu, M. Hofmarcher, V. P. Patil, M. Dorfer, P. M. Blies, J. Brandstetter, J. A. Arjona-Medina, S. Hochreiter, XAI and Strategy Extraction via Reward Redistribution, Springer International Publishing, 2022, pp. 177–205.

[23] B. Beyret, A. Shafti, A. Faisal, Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS, 2019, p. 5014–5019.

[24] S. Miryoosefi, K. Brantley, H. Daume III, M. Dudik, R. E. Schapire, Reinforcement learning with convex constraints, in: Advances in Neural Information Processing Systems, volume 32, 2019, pp. 1–10.

[25] Y. Chow, M. Ghavamzadeh, L. Janson, M. Pavone, Risk-constrained reinforcement learning with percentile risk criteria, Journal of Machine Learning Research 18 (2015) 1–51.

[26] A. Aubret, L. Matignon, S. Hassas, A survey on intrinsic motivation in reinforcement learning, arXiv 1908.06976 (2019).

[27] C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, A survey of preference-based reinforcement learning methods, Journal of Machine Learning Research 18 (2017) 1–46.

[28] A. Andres, E. Villar-Rodriguez, J. D. Ser, Collaborative training of heterogeneous reinforcement learning agents in environments with sparse rewards: What and when to share?, arXiv 2202.12174 (2022).

[29] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, M. Geist, Leverage the average: An analysis of KL regularization in reinforcement learning, in: Advances in Neural Information Processing Systems (NIPS), volume 33, Curran Associates, 2020, pp. 12163–12174.

[30] A. Galashov, S. Jayakumar, L. Hasenclever, D. Tirumala, J. Schwarz, G. Desjardins, W. M. Czarnecki, Y. W. Teh, R. Pascanu, N. Heess, Information asymmetry in KL-regularized RL, in: International Conference on Learning Representations (ICLR), New Orleans, Louisiana, United States, 2019, pp. 1–25.

[31] T. Haarnoja, H. Tang, P. Abbeel, S. Levine, Reinforcement learning with deep energy-based policies, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1352–1361.

[32] M. Persiani, T. Hellström, Policy regularization for legible behavior, arXiv 2203.04303 (2022) 1–16.

[33] L. Rabiner, B. Juang, An introduction to hidden markov models, IEEE ASSP Magazine 3 (1986) 4–16.

[34] F. Yang, S. Balakrishnan, M. J. Wainwright, Statistical and computational guarantees for the baum-welch algorithm, Journal of Machine Learning Research 18 (2017) 1–53.

[35] C. Maree, C. W. Omlin, Reinforcement learning with intrinsic affinity for personalized prosperity management, arXiv 2204.09218 (2022) 1–12.

[36] Knight Frank Company, Knight Frank luxury investment index, 2022. https://www.knightfrank.com/wealthreport/luxury-investment-trends-predictions/, Accessed on 2022-05-27.

[37] T. T.-L. Chong, W.-K. Ng, V. K.-S. Liew, Revisiting the performance of MACD and RSI oscillators, Journal of Risk and Financial Management 7 (2014) 1–12.